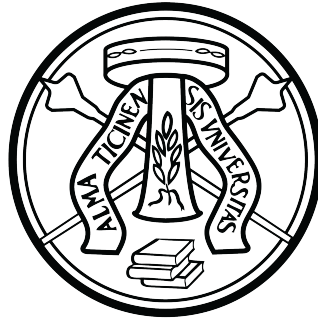




**UNIVERSITÀ
DI PAVIA**

UNIVERSITÀ DEGLI STUDI DI PAVIA

DIPARTIMENTO DI STUDI UMANISTICI

Corso di Laurea Magistrale in

Linguistica Teorica, Applicata e delle Lingue Moderne — LM-39

**Valutazione della Sensibilità Culturale dei LLM
sui Proverbi Italo-Romanzi: Risorse e Benchmark**

RELATRICE

Prof.ssa Elisabetta Ježek

CORRELATRICE

Dott.ssa Sara Tonelli

Tesi di Laurea Magistrale di

Gabriele Errico

Matricola n. [554260]

Anno Accademico 2025/2026

*for those who go the extra mile
for those who put their hearts and souls into it
for those who give and receive less back
for those who are hungry but keep their mouths shut
for those who are held down*

when life feels like a scale, you are the guides of my broken dark roads

*for those whom travel is uncertain
for those who are about to start something new
for those who already embarked on The ferryman's boat
when life feels heavy as a lily's petal flying away from its stem*

*may you be well
may you be safe
may you be healthy
may you be at ease*

*may you be well
may you be safe
may you be healthy
may you be at ease*

Indice

Elenco delle figure	9
Elenco delle tabelle	11
Note terminologiche	15
Introduzione	17
1 Quadro teorico e lavori correlati	23
1.1 Il proverbio: definizione e proprietà linguistiche	24
1.2 Dataset e risorse per i proverbi nel NLP	30
1.2.1 Copertura di lingue ad alta risorsa	30
1.2.2 Copertura di lingue a bassa risorsa	30
1.3 Valutazione dei LLM sui proverbi	32
1.4 La dimensione culturale nel NLP	34
1.4.1 Framework di misurazione culturale	34
1.4.2 Bias e rappresentazione	35
1.4.3 Cultural prompting e i suoi limiti	37
1.4.4 Consapevolezza culturale e task sensibili al contesto	38
1.5 Posizionamento del presente lavoro	41

2	I dati	45
2.1	Raccolta dei dati	45
2.1.1	Criteri di inclusione	48
2.1.2	Schema del dataset	50
2.2	Arricchimento della risorsa	56
2.2.1	Correlazioni paremiologiche	58
2.2.2	Creazione degli scenari	60
2.3	Analisi dei dati	65
2.3.1	Distribuzione geografica e sottoinsieme di valutazione	67
3	Esperimenti	69
3.1	Impostazione degli esperimenti	69
3.1.1	Il task	69
3.1.2	I modelli	70
3.1.3	I prompt	73
3.1.4	Le condizioni geografiche	75
3.1.5	Le metriche di valutazione	75
3.2	Presentazione dei risultati	76
3.2.1	Risultati complessivi	76
3.2.2	Analisi dei pattern	77
3.3	Analisi degli errori	79
3.3.1	Scenari più facili e più difficili	80
3.3.2	Effetto della condizione geografica	82
3.3.3	Analisi per modello	82

4	Discussione	85
4.1	Natura del task	86
4.1.1	Baseline casuale	86
4.1.2	Cautele interpretative	87
4.1.3	Il rischio dell'antropomorfizzazione	88
4.2	Scala parametrica e performance	89
4.2.1	Confronto tra famiglie di modelli	90
4.2.2	Quadro riassuntivo di tutti i modelli	91
4.3	I modelli specializzati sull'italiano	92
4.4	Il segnale geografico: utile ma non sufficiente	94
4.5	Proverbi che hanno viaggiato: memorizzazione vs. sensibilità culturale	96
4.6	Caso di studio: il gruppo VINO → LATTE	98
5	Conclusioni	101
5.1	Sintesi del contributo	101
5.2	Limitazioni	103
5.2.1	Limitazioni della risorsa	103
5.2.2	Limitazioni degli esperimenti	104
5.3	Prospettive future	106
	Appendice	111
A.	Fonti WikiQuote	111
B.	Regioni W-close per DiPRomIt-scenarios	113

C. Test di Wilcoxon signed-rank: assunzioni e calcolo	114
D. Codice per l'elaborazione dei risultati e la riproducibilità . . .	117
E. Note tecniche sulla curatela del dataset	118
Bibliografia	120
Sitografia	137

Elenco delle figure

- 2.1 Pipeline di raccolta e arricchimento dei dati di DiPRomIt e di costruzione di DiPRomIt-scenarios. I riquadri blu: fogli di lavoro prodotti in ciascuna fase (proverbi, scenari); viola: fasi di elaborazione manuale (revisione, creazione degli scenari); il rombo giallo: punto di decisione basato sui criteri di inclusione; verde: fasi di arricchimento e ristrutturazione dei metadati; rosso: elementi esclusi dalla risorsa. Sotto la linea tratteggiata, le direttive sperimentali (Capitolo 3) sono riportate a titolo di contestualizzazione. 49
- 2.2 Grafico a barre orizzontali affiancate che rappresenta la distribuzione delle istanze in ogni regione presente in DiPRomIt-scenarios (N = 147 scenari), in blu, contrapposte a quelle dell'intero DiPRomIt, in rosso. 68
- 3.1 Top- k ($k = 15$) scenari globalmente più facili (*sinistra*) e più difficili (*destra*) per accuracy media aggregata su tutti i modelli, prompt, seed e condizioni geografiche. 80

3.2	Top- k ($k = 15$) scenari più facili (<i>sinistra</i>) e più difficili (<i>destra</i>) disaggregati per condizione geografica (<i>with-geo</i> ✓ e <i>without-geo</i> ✗).	82
3.3	Heatmap che mostra l'accuracy media per modello sui top- k ($k = 15$) scenari più facili (<i>sinistra</i>) e più difficili (<i>destra</i>), aggregata di tutte le condizioni sperimentali.	83

Elenco delle tabelle

1.1	Rassegna delle risorse esistenti per i proverbi nel NLP in ordine di pubblicazione, dal meno recente (sopra) al più recente (sotto). Dial.: la risorsa copre varietà dialettali o non-standard. Uso cont.: il compito richiede l'impiego del proverbio in modo contestuale, non il semplice recupero mnemonico o il completamento. #P: numero di proverbi (o istanze) censiti nella risorsa. n.d.: dato non riportato nella fonte consultata. Icone: 📖 risorsa lessicale/database; 🗣️ narrativa crowdsourced; 🎓 benchmark di valutazione per LLM; 📍 contiene informazione geografica.	43
2.1	Schema del dataset. Le righe evidenziate, <i>Traduzione</i> e <i>Significato</i> , indicano che almeno uno dei due deve essere necessariamente presente. I campi geografici (<i>Regione</i> , <i>Provincia</i> , <i>Città</i>) sono compilati solo quando l'informazione risulta certa.	51
2.2	Esempio di voci in relazione di <i>Alternativa</i>	53
2.3	Esempio di voci in relazione di <i>Variante</i>	54

2.4	Esempio di voce con corrispondenza italiana nota.	56
2.5	Due istanze di proverbi in DiPRomIt con i relativi metadati.	58
2.6	Gruppo di varianti VINO → LATTE. 6 proverbi	59
2.7	Gruppo di varianti LA FAME FA USCIRE IL LUPO. 2 proverbi	59
2.8	Esempio di istanza di DiPRomIt-scenarios (versione <i>with-geo</i> : l'informazione geografica è <u>sottolineata</u>).	60
2.9	Statistiche riassuntive di DiPRomIt per regione. <i>P</i> : istanze di proverbi (#: numero; %: percentuale sul totale); <i>T_{all}</i> : token totali; <i>T_{uniq}</i> : token unici; <i>T/P</i> : media token per proverbio; <i>L</i> : traduzione letterale (# e % dei proverbi con traduzione); <i>M</i> : significato (# e % dei proverbi con significato); <i>S</i> : numero di fonti.	66
3.1	Modelli utilizzati negli esperimenti, con rispettivi identificatori Hugging Face, numero di parametri e origine.	72

- 3.2 Risultati di DiPRomIt-scenarios per tutti i modelli con (✓) e senza (✗) informazione geografica. Accuracy media corredata di deviazione standard su 15 esperimenti per modello per condizione ($147 \times 15 = 2.205$ istanze totali processate). In grassetto i valori migliori assoluti per fascia parametrica. I valori sottolineati indicano l'accuracy nella condizione *with-geo*, per facilitarne il confronto con *without-geo*. Tra parentesi si possono notare i valori assoluti per tipologia di errore. 77
- 4.1 Guadagno di accuracy tra modello minore e maggiore, a parità di famiglia (variante *with-geo*). La riga evidenziata in giallo rappresenta la famiglia con il guadagno minore tra versione piccola e versione grande. 90
- 4.2 Confronto inter-famiglia 7–9B. In giallo il modello in prima posizione. 91
- 4.3 Classifica dei modelli di dimensioni ridotte (7–9B), condizione *with-geo*. In giallo il modello in prima posizione. 91
- 4.4 Classifica dei modelli di dimensioni maggiori (27–72B), condizione *with-geo*. In giallo il modello in prima posizione. 91

1	Elenco degli URL WikiQuote utilizzati per la raccolta di DiPRomIt. Regione: regione amministrativa di riferimento per i proverbi dell'URL. Informazioni geografiche di dettaglio: province e comuni rappresentati (<i>Generica:</i> nessuna informazione geografica di dettaglio disponibile; *: l'informazione geografica corrisponde a un'area che può essere più grande o più piccola di una provincia. In blu: l'informazione geografica in WikiQuote era scritta nel testo del proverbio ed è stata estratta e aggiunta come metadato in DiPRomIt).	112
2	Insieme delle regioni da cui viene campionato il proverbio W_{close} per ciascuna regione in DiPRomIt-scenarios.	113
3	Accuracy nelle due condizioni geografiche per i dieci modelli, differenza Δ e rango assegnato a $ \Delta $ (nessun pareggio, nessuna differenza nulla).	115
4	Script di analisi con una breve spiegazione della funzione e i relativi riferimenti nella tesi.	117

Note terminologiche

Le convenzioni qui elencate sono pensate come riferimento per la lettura dei capitoli successivi. Nel presente lavoro i prestiti dall'inglese non adattati (LLM, benchmark, dataset, prompt, embedding, token, tra i tanti) sono trattati come invariabili: il numero è espresso unicamente tramite l'articolo (il benchmark / i benchmark), mai tramite il suffisso plurale inglese -s. Le sigle seguono la stessa regola (i LLM, non gli LLMs). Fanno eccezione i nomi propri delle due risorse create in questo lavoro, DiPRomIt e DiPRomIt-scenarios).

Di seguito vengono specificate alcune convenzioni tipografiche utilizzate per compilare il seguente lavoro:

- – un trattino, anche denominato *hyphen*, per le parole composte (es. sub-standard);
- -- due trattini, anche chiamati en dash (—), per range numerici (es. 7–9, 2025–2026);
- --- tre trattini, em dash (—), per incisi/parentetiche nel discorso (es. Natural Language Processing — NLP).

-
- l'abbreviazione pp sta per *punti percentuali*, cioè la differenza aritmetica tra due valori percentuali (es. da 40% a 44% ci sono +4 pp).
 - le virgolette caporali: «»; usate per citazioni dirette (prompt, esempi) e per termini o espressioni usati in senso enfatico o cauto (es. «comprensione»).
 - le virgolette doppie: «»; riservate a eventuali citazioni di secondo livello, cioè annidate dentro una citazione già racchiusa tra «».
 - il corsivo è riservato ai prestiti e ai termini tecnici in una lingua diversa dall'italiano (es. *task*, *zero-shot*, *self-attention*, *embedding*); ma anche (es. *quantizzazione*, *neoplurilinguismo*); i proverbi (es. «*Acqua ca duna...*»); le etichette (i)/(ii)/(iii); e i titoli di opere citate;
 - il carattere monospace è utilizzato per i nomi dei modelli (es. Qwen-2.5-72B) e maiuscoletto (come questo) per i nomi delle risorse create in questo lavoro (DiPRomIt, DiPRomIt-scenarios).
 - la virgola è usata come separatore decimale nei numeri e nelle percentuali riportati. Esse sono seguite da un massimo di due cifre decimali (es. 66,2%); il punto è invece riservato ai valori grezzi di accuracy in scala 0–1 (es. 0.4857), riflettendo la notazione usata nelle tabelle dei risultati. Fanno eccezione i valori già di per sé molto piccoli, per i quali un arrotondamento a due cifre altererebbe sensibilmente la precisione (es. Appendice C, +0,0285).

Introduzione

I proverbi fungono da strumenti linguistici tramandati di generazione in generazione per veicolare valori, norme, credenze e visioni del mondo. Fondati sull'esperienza condivisa sono espressioni culturali convenzionalizzate che condensano in forma breve e fissa la saggezza collettiva di una comunità. Poiché si collocano tra lingua, cultura ed esperienze umane, essi costituiscono un oggetto di studio privilegiato sia dal punto di vista linguistico che da quello computazionale, tanto più se si considera la loro dimensione locale, strettamente legata ai dialetti.

I recenti progressi nel campo dell'elaborazione del linguaggio naturale (Natural Language Processing — NLP), resi possibili dall'introduzione dei modelli linguistici di grandi dimensioni (Large Language Model — LLM), hanno prodotto notevoli risultati su un'ampia gamma di task e lingue. Tuttavia, vi è evidenza crescente che i modelli di NLP attuali tendano ad appiattire la diversità culturale e linguistica (Bender et al., 2021). Pertanto sembra mancare un elemento fondamentale: i modelli «scompongono» un problema in minuscole parti e «propongono» alternative plausibili, ma ciò non garantisce una sensibilità, o meglio, una «consapevolezza» completa, di quello che «producono». Manca l'empatia, un'accortezza caratteristica

degli umani.

La sensibilità culturale è quindi uno degli obiettivi del sottocampo del NLP culturale, che si configura pertanto come un parametro fondante di un NLP realmente inclusivo (Liu et al., 2025; Pawar et al., 2025a). Ciò include andare oltre le lingue istituzionalizzate e standardizzate, come le varietà considerate non standard¹, ad esempio i dialetti e le varietà regionali. Rappresentarle diventa quindi necessario, eppure, il NLP attuale presenta ancora marcate disparità di prestazione tra varietà standard e quelle considerate non standard (Faisal et al., 2024). Valutare i LLM su espressioni linguistiche dialettali e culturalmente marcate come i proverbi diventa quindi imprescindibile per comprendere se questi modelli catturino effettivamente, e sfruttino, la conoscenza culturale locale.

Questa introduzione non sarebbe completa senza trattare, anche solo superficialmente, termini come «comprensione» e «awareness», spesso affiancati alla terminologia NLP. Come osservava Cristianini (2023), che citava modelli antecedenti a quelli di frontiera, come AlphaGo e GPT-3 (oggi osserviamo una loro rapida evoluzione), relazioni statistiche, dall'autore presentate brillantemente come «scorciatoie», avrebbero rimpiazzato il normale ragionamento dietro la «comprensione» dei concetti. In breve, l'intelligenza artificiale (Artificial Intelligence — AI) non ha bisogno di «ragionamento», perché «comprende» i concetti in maniera diversa da

¹ Con *non standard* si intende qualsiasi varietà priva di attestazione standard codificata, cioè senza implicazioni valutative. La nozione si differenzia da quella di *sub-standard*, che designa varietà stigmatizzate per ragioni di prestigio diastratico, come per l'italiano popolare (cfr. Berruto, 2014).

quella umana, in questa opera intesa appunto come una scorciatoia. Che il fenomeno della comprensione risulti simulato, e di conseguenza l'illusione che il modello abbia «ragionato» è quindi altamente probabile; di converso, è modestamente probabile e insufficientemente verificabile, ammettere che basti una grossa quantità di dati per raggiungere una vera e propria «comprensione». Per quanto auspicabile per un essere umano, è noto che quest'ultimo ha a disposizione svariati altri strumenti ben più complessi (come i sensi e le esperienze, tra i tanti) per poter comprendere efficacemente un concetto. In questa tesi si prova a dare un'interpretazione alla luce dei risultati sperimentali.

Da questo quadro introduttivo emergono quattro domande di ricerca che guidano il presente lavoro:

RQ1 I modelli linguistici abbinano correttamente i proverbi allo scenario appropriato, oppure la possibile incompetenza è riconducibile a una causa, in questo caso a meccanismi di memorizzazione?

RQ2 L'inclusione esplicita di informazioni sulla provenienza geografica nel prompt fornisce ai modelli un segnale utile per l'abbinamento corretto del proverbio?

RQ3 Quali pattern emergono dall'analisi degli errori dei modelli, in particolare rispetto a quanto appena descritto?

RQ4 I modelli mostrano una forma di conoscenza dei proverbi emergente, paragonabile a quella della gente comune?

Il presente lavoro si articola attorno a due contributi principali. Il primo consiste nella raccolta manuale di proverbi dialettali italo-romanzi e nella loro sistematica strutturazione in DiPRomIt (Dialectal Proverbs written in local Romance languages of Italy), il primo dataset di proverbi scritti in varietà di lingue romanze parlate in Italia, uno dei paesi linguisticamente più eterogenei² d’Europa (Ramponi, 2024).

DiPRomIt comprende 1.310 proverbi curati, arricchiti di metadati geografici, traduzione letterale, significato, fonte originale e correlazioni paremiologiche inter-regionali.

Il secondo contributo consiste nella creazione manuale di DiPRomIt-scenarios, un benchmark composto manualmente da 147 scenari corredati di contesto, progettato per valutare la sensibilità culturale *contestuale* dei LLM. Su DiPRomIt-scenarios viene valutato un insieme di 10 LLM open-weight³ in condizioni sperimentali variabili, testando tra l’altro se l’inclusione di informazioni geografiche nel prompt fornisca segnali utili per l’abbinamento corretto del proverbio.

I risultati mostrano che la sensibilità culturale sul task è ancora lontana dall’essere raggiunta dai modelli attuali: il miglior modello abbina correttamente circa due terzi dei proverbi del benchmark. Emerge inoltre che i proverbi dialettali privi di equivalente italiano, ovvero quelli che non

² Il concetto di «eterogeneità linguistica» si intende qui come la componente storica del plurilinguismo italiano, costituita da parlate italo-romanze e dalle minoranze linguistiche autoctone, distinta invece dal cosiddetto *neoplurilinguismo* italiano (cfr. Fiorentini, 2022).

³ Con *open-weight* (o «open-source») si intendono quei modelli linguistici i cui pesi interni, cioè i parametri addestrati (*weights*), sono pubblicamente (*open*) accessibili, scaricabili e interrogabili, anche localmente. Si differenziano dai modelli closed-source o proprietari, che per definizione sono accessibili solo tramite chiamate API; per una trattazione più dettagliata cfr. Capitolo 3, §3.1.2.

sono circolati tra le lingue, risultano più difficili da abbinare allo scenario corretto. Entrambi i dataset, DiPRomIt e DiPRomIt-scenarios, sono resi pubblicamente disponibili alla comunità di ricerca.

Capitolo 1

Quadro teorico e lavori correlati

Il primo capitolo si articola in cinque sezioni. La Sezione 1.1 introduce il proverbio: oggetto di studio linguistico-computazionale della presente tesi, passando dalla sua natura non compositazionale ed esplicitando le convenzioni terminologiche utilizzate dalla tradizione paremiologica. Vengono proposte, a titolo inclusivo, anche le sfide cognitive poste dalla elaborazione dei proverbi. La Sezione 1.2 presenta una rassegna aggiornata e rappresentativa delle risorse esistenti per i proverbi nel NLP, con particolare attenzione alla distinta rappresentazione tra lingue ad alta e a bassa risorsa. La Sezione 1.3 esamina lo stato dell'arte nella valutazione dei LLM su task legati ai proverbi. La Sezione 1.4 discute il ruolo dei framework culturali nella ricerca del NLP culturale, dai sondaggi psicometrici alla presentazione dei principali *position paper*.

La Sezione 1.5 illustra infine come il presente lavoro si collochi e si differenzi rispetto alla letteratura esaminata.

1.1 Il proverbio: definizione e proprietà linguistiche

Dare una definizione univoca di proverbio risulterebbe alquanto difficile, per questo motivo questa sezione ripercorre diverse definizioni operative proposte in letteratura, muovendosi da quella più generale fino alla formulazione elaborata dalla tradizione geoparemiologica italiana, proposta come sintesi. Per la stessa ragione, la trattazione integra prospettive teoriche diverse senza privilegiarne una a priori: ciò riflette un tentativo di articolazione complementare piuttosto che un'adesione esclusiva a una singola scuola.

Un proverbio è un'espressione tradizionale di saggezza popolare che trasmette verità percepite, fondate sul senso comune o sull'esperienza collettiva locale. In quanto dispositivo culturale, funge da strumento linguistico tramandato di generazione in generazione per veicolare valori, norme e visioni del mondo (Mieder, 2004)¹. La definizione lessicografica canonica lo descrive come segue:

«Breve motto, di larga diffusione e antica tradizione, che esprime, in forma stringata e incisiva, un pensiero o, più spesso, una norma desunti dall'esperienza»²

Un'ulteriore definizione di proverbio ci viene data dalla sociolinguistica a proposito di etnotesto: i proverbi rientrano infatti tra i «testi fissati dal-

¹ Per una rassegna più ampia delle definizioni di proverbio proposte nella tradizione paremiologica, cfr. Mieder (2004).

² Treccani, voce *Proverbio* (consultato il 12 maggio 2026). Essendo della famiglia lessicale dei *motti*, può essere utile fare riferimento anche alla voce *Motto* (consultato il 12 maggio 2026).

la tradizione» che, insieme ai testi liberi, come ricordi autobiografici e testimonianze di usi, costituiscono quei «testi orali che rappresentano un'espressione autonoma della cultura di una comunità linguistica»³ (Beccaria, 2004).

Poiché i proverbi qui esaminati concernono varietà dialettali, è necessaria una precisazione terminologica. Per dialetti italo-romanzi si intendono, in accordo con la tradizione della linguistica romanza, sistemi linguistici distinti dall'italiano e dotati di storia autonoma: si tratta di dialetti primari, varietà sorelle dell'italiano che si sono sviluppate in parallelo a partire dai volgari romanzi medievali, e non, ad esempio, come sottovarietà dell'italiano standard (cfr. Berruto & Cerruti, 2015, cap. 3).

Applicando la distinzione di Kloss (1978) tra *Abstandsprache* (lingua per distanziamento, riconoscibile come sistema a sé in virtù delle proprie caratteristiche strutturali) e *Ausbausprache* (lingua per elaborazione, dotata di standard scritto consolidato), i dialetti italo-romanzi si configurano come sistemi con un sufficiente grado di distanziamento strutturale ma con un minore grado di elaborazione rispetto all'italiano.

In questo senso, la situazione italiana è stata descritta come una lingua *cum dialectis*, nella quale «dialetto e lingua nazionale vanno considerati sistemi separati per ragioni sia di autonomia storica [...] che di differenza strutturale, per cui tra la lingua nazionale e alcuni dialetti vige una distanza non inferiore dal punto di vista lessicale, fonetico e morfologico a quella

³ Treccani, voce *La lessicologia dialettale*.

esistente tra l'italiano e altre lingue neolatine» (Cerruti & Regis, 2005, p. 180).

L'etichetta dialetto, quando citata in questo lavoro, è quindi adottata convenzionalmente: riflette una categorizzazione storico-sociale e non un giudizio sul loro statuto di sistemi linguistici indipendenti. È in queste varietà che si radica il proverbio oggetto del presente lavoro.

Nell'ambito della tradizione paremiologica italiana, Temistocle Franceschi, fondatore della Scuola Geoparemiologica Italiana⁴, descrive il proverbio come un enunciato di natura fondamentalmente retorica: una formula sintetica e allusiva che, anziché designare direttamente il proprio referente, rimanda per analogia a un insieme semantico diverso da quello apparente, esprimendo così in modo indiretto un parere o un consiglio (Franceschi, 2004). Questa natura allusiva è per sé polisemica, che Franceschi definisce *valore paremiaco*⁵. All'interno di questa cornice, Franceschi distingue funzionalmente i *detti didattici* — tipici dell'ambito meteorologico-calendario o igienico — dai *detti paremiaci*, che veicolano invece un'opinione indiretta condivisa, nonché analogica (Franceschi, 2004). È proprio questa seconda categoria, per la sua natura allusiva e non immediatamente decifrabile, a porre le sfide interpretative discusse di seguito.

Definita la guida dell'oggetto di studio, è utile introdurre una dimensione fondamentale per spiegare la natura dei proverbi. Dal punto di vista

⁴ La geoparemiologia è la disciplina che studia la variazione geografica dei proverbi all'interno di un territorio.

⁵ Franceschi evita deliberatamente il termine «significato»: un'espressione che presupporrebbe un'interpretazione univoca, da trovare, laddove il proverbio ammette per sua natura una gamma di valori possibili più ampia, non un singolo valore (Franceschi, 2007).

semantico, il principio di composizionalità, cardine della semantica formale a partire dal lavoro di Frege (1892) e sistematizzato da Partee (1995), stabilisce che il significato di un'espressione complessa sia funzione del significato delle sue parti costitutive e delle regole sintattiche che le combinano. Tale principio, tuttavia, incontra un limite sistematico nel momento in cui si considerano espressioni il cui significato non risulta dalla semplice somma delle parti: è il caso, in particolare, delle espressioni pluri-lessicali (MWE, *Multi Word Expressions*), tra cui i proverbi rappresentano un caso paradigmatico.

La letteratura converge su tre asserzioni distinte (Culicover et al., 2017; Dickins, 2020; Lemghari, 2021): (i) la sintassi nel proverbio risulta solo in parte rispettata; (ii) la composizionalità del significato risulta parziale; (iii) l'interpretazione del significato annesso al proverbio risulterebbe di natura pressoché pragmatica, che richiederebbe un'inferenza analogica indiretta tra l'enunciato proverbiale e il contesto in cui viene evocato.

A titolo illustrativo: il significato di *Una mela al giorno leva il medico di torno* non è la somma dei significati delle singole parole, bensì la loro commistione profonda (Azime et al., 2025). Il suggerimento sottostante è infatti quello di intraprendere un'azione preventiva per evitare una conseguenza spiacevole, non che l'ingestione di una mela privi il medico del suo lavoro. Nella stessa prospettiva, la paremiologia italiana descrive lo stesso fenomeno attraverso la nozione di *macrolemma*: un'intera espressione proverbiale trattata come un'unica unità lessicale, non scomponibile nei propri

costituenti⁶.

Comprendere appieno i proverbi e saperli utilizzare nel loro contesto non è un'operazione banale. Questa solo apparentemente semplice operazione prevede delle sfide cognitive consistenti: innanzitutto conoscere solo il significato del proverbio non è sufficiente, bisogna selezionarne il senso più appropriato, letterale o connotativo, e integrarlo nel discorso (Katz & Ferretti, 2001; Temple & Honeck, 1999).

La *graded salience hypothesis* propone tre principi fondamentali che si troverebbero alla base del processo di decodifica, attraverso il quale la nostra mente deve lavorare in modo da non produrre un enunciato fuori contesto: (i) *conventionality (conformity to tradition* — convenzionalità, affidabilità, o aderenza alla tradizione), (ii) *frequency* (frequenza d'uso, cioè in termini di distribuzione) e (iii) *familiarity* (familiarità, o meglio, grado di certezza). Questi tre fattori guidano la risoluzione della competizione cognitiva tra interpretazione letterale e figurata (Giora, 1997). Studi empirici hanno confermato che alcuni proverbi hanno un significato più letterale che figurato di altri, e questi risultano anche più facilmente annotabili mentalmente, rispetto a quelli figurativi (Katz & Ferretti, 2001).

Dal punto di vista semantico-cognitivo, il significato del proverbio è analizzato come struttura concettuale il cui «concetto base» deve essere esemplificato da un argomento contestualizzato⁷. Ciò rafforza la teoria del

⁶ Nella terminologia di Franceschi, l'insieme dei macrolemmi di natura proverbiale costituisce il *codice paremiaco*, a sua volta sottoinsieme del più ampio *codice retorico*, che comprende l'intera espressione figurata di una lingua (Franceschi, 2007).

⁷ Un meccanismo concettualmente analogo, sul piano della semantica lessicale anziché fra-seologica, è descritto dalla nozione di *enriched compositionality* (Pustejovsky, 2002): attraverso

trasferimento analogico (Lemghari, 2021).

Non mancano infine tentativi di formalizzazione, come la *partial compositionality* (Socolof, 2024). Come non mancano quelli di ontologizzazione attraverso risorse strutturate, come ProverbNet, progettato per analizzare il passaggio dal significato letterale a quello inteso (Pimpalgaonkar et al., 2021).

Le nozioni sin qui discusse confluiscono nella definizione conclusiva proposta dalla Scuola Geoparemiologica Italiana, che vale la pena riportare per intero:

«Possiamo a questo punto tracciare una definizione del proverbio come breve e concisa sentenza allologica (ossia, intesa a comunicare altro da sé) di struttura euritmica e oppositoria, che in un determinato idioma è convenzionalmente usata in riferimento allusivo ad altro insieme semantico con cui viene analogicamente correlata, così da esprimere in modo indiretto, sintetico e retoricamente efficace un parere, un commento, un consiglio. Ovvero: formula polisemica brachilogica allologica analogica, che, sotto l'apparenza di un'informazione specifica, svolge la funzione di comunicare succintamente e per via indiretta un'opinione sui più varî aspetti della vita umana. Et similia.» (Franceschi, 2007)

processi come *co-composition* e *type coercion*, il significato di un'espressione si arricchisce dinamicamente in base al contesto.

1.2 Dataset e risorse per i proverbi nel NLP

La rassegna della letteratura si articola lungo due dimensioni intersecanti (Oh et al., 2025): (i) la rappresentazione delle varietà linguistiche nelle risorse esistenti; (ii) la valutazione dei modelli su task che coinvolgono i proverbi.

1.2.1 Copertura di lingue ad alta risorsa

La maggior parte delle risorse esistenti copre lingue standardizzate ad alta risorsa⁸. Il dataset Maps (Liu et al., 2024) comprende circa 300–500 proverbi per ciascuna di sei lingue (inglese, cinese mandarino, tedesco, russo, bengalese e indonesiano), successivamente adattato per la traduzione automatica (M. Wang, Pham et al., 2025). Il dataset ProMetheus (Özbal et al., 2016) contiene 1.054 proverbi in inglese e italiano annotati con metafore. Ghosh e Srivastava (2022) propongono ePIC, un insieme di coppie proverbio-narrativa in inglese progettato per la valutazione del recupero di informazioni e della generazione di testi.

1.2.2 Copertura di lingue a bassa risorsa

Le risorse contenenti proverbi dialettali rimangono estremamente scarse.

Tra le eccezioni più rilevanti, Magdy et al. (2025) introducono Jawaher, un

⁸ Con *lingua ad alta* (o *bassa*) *risorsa* si intende una lingua per cui sono disponibili grandi (o scarse) quantità di dati. La dimensione di scarsità/abbondanza rappresentativa non si limita però ai soli dati grezzi, bensì a corpora annotati e modelli pre-addestrati, e questi sono solo alcuni tratti di questa sproporzione che automaticamente sfavorisce le lingue a bassa risorsa (Dong et al., 2024; Mitkov, 2014).

dataset multi-dialettale che copre 20 varietà della lingua araba. In aggiunta, un atlante di proverbi greci con varianti locali è stato costruito per consentire l'esplorazione computazionale e l'attribuzione geografica dei proverbi (Pavlopoulos et al., 2024).

Anche per le varietà italo-romanze la copertura è bassa. Le risorse dedicate alle varietà romanze parlate in Italia comprendono innanzitutto gli atlanti linguistici, quali l'*Atlante Linguistico Italiano* (ALI) (Bartoli et al., 1995), l'*Atlante Linguistico ed Etnografico dell'Italia e della Svizzera Meridionale* (AIS) (Jaberg et al., 1987) e l'*Atlante Lessicale Toscano* (ALT) (Giacomelli, 2000), oltre ad altri dataset non specificamente incentrati sui proverbi. Tra questi, Ramponi e Casula (2023a) introducono DiatopIt, un corpus per lo studio della variazione diatopica in Italia, cui è associato lo shared task GeoLingIt (Ramponi & Casula, 2023b); Calderaro e Monti (2026) propongono un corpus di poesia estemporanea in sardo logudorese; Haberland et al. (2024) introducono un corpus parallelo genovese-italiano per la traduzione automatica; Cappello et al. (2025) sviluppano un treebank parallelo siciliano-italiano; Sprugnoli et al. (2025) creano un lemmario siciliano e un glossario bilingue siciliano-italiano. Per l'italiano, Mensa et al. (2025) introducono ProverbIT, un benchmark di 100 domande a scelta multipla, basate su proverbi italiani, utilizzato per testare la capacità dei LLM di «comprensione» del significato.

Lo studio mostra come i modelli ricorrano a euristiche superficiali nel task di completare delle parti mancanti di proverbi e faticino a riconoscere

l'assenza della risposta corretta, quando questa *non* è tra le opzioni proposte. Il benchmark è stato successivamente adottato nell'iniziativa di valutazione CALAMITA (Mensa et al., 2026).

Alcune risorse si concentrano su lingue non europee a bassa risorsa, con l'obiettivo di valutare la capacità dei LLM di interpretare espressioni culturalmente marcate: è il caso del nepalese (Thapa et al., 2025), delle lingue etiopi (Azime et al., 2025), del sanscrito e del marathi (Pimpalgaonkar et al., 2021).

1.3 Valutazione dei LLM sui proverbi

Il linguaggio figurato è considerato un banco di prova privilegiato per i modelli linguistici proprio perché il suo significato non è recuperabile per composizione (Almeida et al., 2025). I proverbi rendono i modelli vulnerabili alle «allucinazioni» (Smith et al., 2023): essi producono risposte plausibili, ma fattualmente errate (Hosseini et al., 2026).

Valutazioni recenti mostrano come i modelli di *embedding*⁹ e di rappresentazione contestuale non siano consistenti nei risultati dei test (Kommers et al., 2025): anche quando sembrano ottenere buoni punteggi, dimostrano di affidarsi a segnali lessicali superficiali, non sufficienti a risolvere le ambiguità (He et al., 2025). Nel benchmark ePIC, il tasso di *retrieval* scende al 6,6% quando ai modelli viene semplicemente richiesto di collegare i

⁹ Seppure gli embedding contestuali superino i modelli statici risolvendo, ad esempio, «casi ambigui di polisemia» (Devlin et al., 2019), essi derivano le proprie rappresentazioni da regolarità distribuzionali. Ciò risulta ancora insufficiente a discriminare un impiego pertinente da uno solamente plausibile.

proverbi alla narrativa corrispondente, senza averla vista in precedenza (Ghosh & Srivastava, 2022).

Nel corso degli anni alcune abilità di recupero delle informazioni sono migliorate, anche in contesti multilingue, tuttavia emergono nuove problematiche: bias culturali nei dati di addestramento inducono i modelli a prediligere determinate culture piuttosto che altre, come le culture anglofone e cinesi, mentre il task di comprensione del senso figurato rimane un collo di bottiglia irrisolto (Liu et al., 2024).

Nell'ambito della traduzione, alcuni studi mostrano che i LLM possono superare i modelli di MT (*Machine Translation*) nell'inferire alcune sfumature idiomatiche, tuttavia le metriche automatiche come BLEU e COMET, utilizzate in questo campo, si sono dimostrate inadeguate per misurare l'appropriatezza culturale (M. Wang, Adel et al., 2025). I modelli faticano inoltre a distinguere le espressioni autentiche da quelle inventate ma plausibili (Mensa et al., 2025).

Sembrerebbe che per ovviare ad alcuni problemi basterebbe inquadrare l'input in un contesto illustrativo specifico, o chiedere il significato del proverbio prima del task. Questi percorsi indurrebbero a favorire l'etichettatura a livello di token; tuttavia, quando si trattano argomenti più astratti, come la comprensione della metaforicità, nella sua accezione più ampia, un'accuratezza soddisfacente rimane lontana dall'essere realizzata (Goren & Strapparava, 2024).

1.4 La dimensione culturale nel NLP

La dimensione culturale della comprensione linguistica dei LLM è oggetto di un filone di ricerca in rapida espansione, alimentato dalla necessità di verificare se questi modelli, addestrati su dati sbilanciati verso alcune culture, siano in grado di riconoscere e applicare correttamente conoscenza culturalmente specifica. La rassegna che segue si articola in quattro momenti: si presentano anzitutto gli strumenti impiegati per misurare l'allineamento culturale dei modelli, dai sondaggi sociologici alle scale psicometriche; si esaminano poi le fonti di bias e sotto-rappresentazione che ne condizionano l'affidabilità, in particolare la sovra-rappresentazione di alcune strutture rispetto ad altre; si discutono le tecniche di *cultural prompting* proposte per correggere tali squilibri e i loro limiti strutturali; si conclude infine con la distinzione, dibattito più recente, tra allineamento culturale e consapevolezza culturale.

1.4.1 Framework di misurazione culturale

La ricerca culturale sui LLM si è consolidata attorno all'uso di sondaggi sociologici, mirati a rilevare atteggiamenti e valori a livello di popolazione, e scale psicometriche, valide su tratti individuali (Pawar et al., 2025a). Lo strumento più utilizzato è il WVS, *World Values Survey*¹⁰, un progetto globale per la raccolta di risposte su questioni sociali, etiche ed economiche,

¹⁰ <https://www.worldvaluessurvey.org/>, consultato il 27 maggio 2025. Cfr. anche World Values Survey Association, 2025.

spesso affiancato dal PGAS, *Pew Global Attitudes Survey*, spesso utilizzato in ricerca NLP nel tentativo di allineare le prospettive umano-macchina.

A parità di utilizzo si trovano le Dimensioni Culturali di Hofstede, sintetizzate nel VSM, *Value Survey Module*: questo test permette di analizzare le risposte dei modelli sulle dimensioni individualismo/collettivismo, rendendo possibile mappare le tendenze soggiacenti alle loro risposte superficiali (Kabir et al., 2025).

La Teoria dei Valori di Schwartz (Schwartz, 1992, 2012) mira a testare i valori intesi come credenze o stati finali desiderabili, ordinati per importanza, che guidano la valutazione di eventi e comportamenti umani (Liu et al., 2025). Studi hanno dimostrato che molti dei suoi elementi sarebbero già incorporati nei dati del WVS, confermando la comparabilità dei due framework (Kaasa & Welzel, 2023). Questa teoria è al centro di progetti innovativi volti a identificare e prevenire rischi derivanti dal disallineamento dei modelli (Yao et al., 2024).

1.4.2 Bias e rappresentazione

La ricerca ha ampiamente documentato la sovra-rappresentazione di strutture WEIRD, cioè la tendenza della ricerca psicologica e delle scienze sociali a basarsi quasi esclusivamente su campioni provenienti da società occidentali (*West*), istruite (*Educated*), industrializzate (*Industrialized*), ricche (*Rich*) e democratiche (*Democratic*) (Henrich et al., 2010), quelle più rappresentate nei dati di addestramento dei LLM (Bulté & Rigouts Terryn, 2026; Masoud

et al., 2025; Oh et al., 2025). Questa asimmetria è visibile già a livello dei dati: alcune regioni, come quelle dell’America Latina, sono rappresentate per meno dello 0,5% dei dati di addestramento (Almeida et al., 2025). Analisi di comunità online come Reddit (Borenstein et al., 2025) confermano che la cultura estratta dal web è sistematicamente sbilanciata.

Oltre alla sotto-rappresentazione quantitativa dei dati, un’ulteriore fonte di distorsione riguarda la misurazione stessa dei valori: i domini legati alla Teoria di Schwartz rappresentano una sfida specifica, poiché, a differenza di concetti istituzionali, i valori sono intrisi di metafore e interpretazioni soggettive che resistono alla mappatura lessicale diretta (Guo et al., 2025). La teoria evidenzia altresì come certi valori siano incompatibili tra culture diverse — si pensi al concetto occidentale di purezza, difficilmente conciliabile con una prospettiva taoista — portando i modelli a errori di valutazione o a preferenze sistematicamente polarizzate.

Vi è infine uno spostamento epistemico in corso nella disciplina: dalla valutazione della cultura come insieme di fatti statici, spesso ridotta a un quiz a scelta multipla, al rilevamento di profili di bias latenti inferibili dalle risposte (Pawar et al., 2025a; Vo et al., 2025; Zhou et al., 2025). Questo approccio a quiz non tiene conto né della natura probabilistica delle domande a scelta multipla, né della complessità dei concetti sottostanti, né del caso in cui la risposta corretta sia assente tra le opzioni (Oh et al., 2025).

L’allineamento culturale, *Cultural alignment* — ovvero la misura in cui le risposte di un modello riflettono i valori e le norme di uno specifico

gruppo umano (AlKhamissi et al., 2024) — rimane superficiale: i modelli conoscono i proverbi a livello di recupero delle informazioni, ma non li impiegano appropriatamente in situazioni che richiedono uno sforzo cognitivo contestuale (Liu et al., 2024; Xu et al., 2025). Nel benchmark CASA (Qiu et al., 2025), nato per misurare la *awareness coverage*, che incentra la sua ricerca su due task *web-based* in lingua inglese (acquisti online e forum di discussione sociale), i risultati degli agenti sono inferiori al 10%, indice di bassa copertura. Le tecniche di *alignment* come RLHF, *Reinforcement Learning from Human Feedback*, e DPO, *Direct Preference Optimization*, hanno migliorato l’adattabilità alle preferenze degli utenti, ma non hanno ancora colmato il divario (Magdy et al., 2025).

1.4.3 Cultural prompting e i suoi limiti

Una linea di ricerca parallela ha esplorato tecniche di prompting per indurre nei modelli un comportamento culturalmente orientato. Nasce così il *cultural prompting* (Tao et al., 2024), che istruisce il modello ad adottare la prospettiva di un parlante appartenente a una specifica cultura o nazione, tipicamente tramite un’istruzione di sistema del tipo «*rispondi come farebbe una persona di X*», dove X è un segnaposto per la cultura o nazione specificata nel prompt (ad es. «...del Giappone»).

Approcci correlati includono il *persona prompting* (Lutz et al., 2025), che assegna al modello un’identità sociodemografica, e l’*anthropological prompting* (AlKhamissi et al., 2024), che nasce per cercare di includere

gruppi sottorappresentati. Quest'ultimo arricchisce l'istruzione di ruolo con informazioni sociodemografiche più granulari rispetto ai precedenti (regione, età, istruzione). Entrambi gli approcci si distinguono dal più semplice *cultural prompting* per il grado di specificità del profilo assegnato al modello, ma condividono la stessa limitazione di fondo: presuppongono che il modello abbia già acquisito durante il pre-addestramento una conoscenza culturale sufficiente da poter essere attivata tramite il prompt (Pawar et al., 2025a).

Questi metodi hanno mostrato benefici misurabili per l'allineamento con dati di sondaggi psicometrici, ma presentano limiti strutturali rilevanti: il comportamento culturale indotto risulterebbe instabile tra contesti diversi e tenderebbe a riflettere stereotipi culturali dominanti piuttosto che la variazione intra-culturale (Pawar et al., 2025a). Soprattutto, il cultural prompting agisce sulle *preferenze espresse*: un modello impara a dichiarare valori culturali coerenti con un gruppo senza saper applicare le espressioni di quel gruppo in situazioni reali.

1.4.4 Consapevolezza culturale e task sensibili al contesto

Come si è potuto evincere accanto al concetto di *allineamento culturale* (*cultural alignment*), la letteratura recente ha elaborato una nozione distinta e più fine: la *consapevolezza culturale* (*cultural awareness*), intesa come la capacità di un sistema di comprendere il contesto in cui viene richiesta l'esecuzione di un task e di interpretare gli elementi sulla base di quel

contesto (Pawar et al., 2025a).

La distinzione è rilevante sul piano metodologico: mentre l'allineamento culturale misura la corrispondenza tra le risposte del modello e i valori di un gruppo — tipicamente tramite sondaggi psicometrici come WVS e Hofstede — la consapevolezza culturale riguarda la capacità di operare correttamente in *task sensibili al contesto* (*context-sensitive tasks*). I proverbi costituiscono un caso paradigmatico di task sensibile al contesto: capacità di abbinare un'espressione culturalmente radicata a una situazione specifica — inferenza che presuppone *ancoraggio culturale* (*cultural grounding*).

Lavori recenti mostrano che né la dimensione della scala parametrica né l'esposizione multilingue fanno emergere spontaneamente «conoscenza» culturale. E che i benchmark esistenti la sottostimano privilegiando task di recupero (Liu et al., 2024; Qiu et al., 2025). Questa critica trova eco in un filone di *position paper* recenti, che convergono nel denunciare i limiti dell'approccio dominante alla valutazione culturale dei LLM.

Zhou et al. (2025) sostengono che il campo manchi di una teoria condivisa di «cultura», affidandosi a proxy grossolani — confini nazionali, copertura limitata alle culture più rappresentate, benchmark statici incapaci di seguirne l'evoluzione — e propongono di attingere alla linguistica socioculturale, suggerendo la nozione di *localizzazione* come cornice più utile della «cultura» tout court. Oh et al. (2025) criticano il «paradigma della triviale», per cui la cultura viene ridotta a fatti statici testabili tramite domande a scelta multipla, e sostengono che gli assunti culturali permeino

anche le impostazioni di valutazione apparentemente neutre, richiamando l'attenzione sul posizionamento di chi valuta.

Alkhamissi et al. (2026) esaminano venti benchmark culturali ed individuando sei problemi metodologici ricorrenti: tra cui il trattare le nazioni come culture e il trascurare la variazione intra-culturale. Costoro propongono di integrare narrazioni e scenari più realistici, valutando i modelli nel contesto anziché in isolamento. Kommers et al. (2025), infine, propongono l'*ermeneutica computazionale*: i sistemi generativi funzionano come da «macchine di contesto» alle prese con situatezza, pluralità e ambiguità interpretativa, e la valutazione dovrebbe quindi essere iterativa; coinvolgere persone oltre alle macchine, e misurare il contesto culturale, mettendo meno enfasi solo sull'output del modello.

Alla critica si affianca una circostanza di fatto propria delle tecniche di prompting culturale discusse in precedenza: finché sarà l'essere umano a dover specificare esplicitamente, tramite il prompt, quale prospettiva culturale il modello debba adottare, non si può parlare propriamente né di allineamento né tantomeno di consapevolezza culturale, poiché entrambe le nozioni presuppongono che sia il sistema, e non l'utente, a riconoscere le circostanze rilevanti. È questa la lacuna che DiPRomIt-scenarios intende principalmente colmare.

1.5 Posizionamento del presente lavoro

Il presente lavoro si differenzia dalla letteratura esistente per due ragioni principali.

In primo luogo, raccogliamo, curiamo e arricchiamo manualmente DiPRomIt, un dataset di proverbi di lingue romanze d'Italia, mettendo a disposizione della comunità una nuova risorsa per l'indagine linguistica. Colmiamo così una lacuna¹¹ nella rappresentazione delle varietà parlate in Italia. Si sottolinea il fatto che il panorama italiano sia tra i più linguisticamente eterogenei d'Europa (Ramponi, 2024). A differenza delle risorse esistenti, che coprono principalmente lingue standardizzate o, nel caso dialettale, varietà non romanze come l'arabo (Magdy et al., 2025), DiPRomIt si concentra esclusivamente sulle varietà italo-romanze.

In secondo luogo, affrontiamo una lacuna altrettanto rilevante nelle pratiche di valutazione dei LLM sui proverbi: i benchmark esistenti testano prevalentemente la capacità di *memorizzare* un proverbio, non di impiegarlo in un contesto appropriato, che è, sosteniamo, il task più adeguato per testare la sensibilità culturale. La conoscenza culturale è tipicamente indistinguibile dal recupero mnemonico, e a nostra conoscenza solo Ghosh e Srivastava (2022) mirano esplicitamente all'uso contestuale dei proverbi, pur limitandosi all'inglese standard.

Con DiPRomIt-scenarios proponiamo una nuova metodologia per valutare la sensibilità culturale dei LLM attraverso la selezione del proverbio

¹¹ Una scarsità di rappresentazione, che si intende, per scopi computazionali.

più appropriato in risposta a scenari contestualizzati. La Tabella che segue 1.1 colloca DiPRomIt e DiPRomIt-scenarios nel panorama delle risorse esistenti riguardanti i proverbi nel NLP. Sono evidenziate le caratteristiche distintive rispetto ai lavori precedenti.

Tabella 1.1: Rassegna delle risorse esistenti per i proverbi nel NLP in ordine di pubblicazione, dal meno recente (sopra) al più recente (sotto). **Dial.:** la risorsa copre varietà dialettali o non-standard. **Uso cont.:** il compito richiede l’impiego del proverbio in modo contestuale, non il semplice recupero mnemonico o il completamento. **#P:** numero di proverbi (o istanze) censiti nella risorsa. n.d.: dato non riportato nella fonte consultata. **Icone:** 📖 risorsa lessicale/database; 🗺️ narrativa crowdsourced; 🏠 benchmark di valutazione per LLM; 📍 contiene informazione geografica.

Lavoro	Lingua/e	Genere e dominio	Compito	#P	Dial.	Uso cont.
<i>Varietà standard</i>						
ProMetheus (Özbal et al., 2016)	en; it	📖 generale	annotazione di metafore	1.054	No	No
ProverbNet (Pimpalgaonkar et al., 2021)	sa; mr	📖 generale (database multilingue)	catalogazione e ricerca	3.456	No	No
ePIC (Ghosh & Srivastava, 2022)	en	🗺️ narrativa (crowdsourcing)	recupero e generazione	n.d.	No	Sì
MAPS (Liu et al., 2024)	en; zh; de; ru; bn; id	🗺️ situazionale/conversazionale	scelta multipla	~300–500 /lingua	No	No
Prone (Thapa et al., 2025)	ne	🏠 generale/culturale	scelta multipla; categorizzazione	2.830	No	No
ProverbEval (Azime et al., 2025)	am; ti; gez; om; en	🏠 generale/culturale, bassa risorsa	scelta multipla; generazione	n.d.	No	No
<i>Varietà dialettali e non-standard</i>						
Towards a Greek Proverb Atlas (Pavlopoulos et al., 2024)	el	📍 proverbi regionali (geocoding)	attribuzione geografica	>100.000 varianti	Sì	No
Jawaher (Magdy et al., 2025)	ar	🏠 generale/culturale	traduzione e spiegazione	10.037	Sì	No
ProverbIT (Mensa et al., 2025)	it	🏠 generale	scelta multipla (compl. + sign.)	100	No	No
DiPRomIt (nostro) (Errico et al., 2026)	varietà italo-romanze	📍 generale/culturale	selezione contestuale (scenari)	1.310 (147 scenari)	Sì	Sì

Capitolo 2

I dati

Il capitolo si articola in tre sezioni. La Sezione 2.1 descrive le modalità di raccolta dei dati, i criteri di inclusione adottati e lo schema della risorsa. La Sezione 2.2 illustra le operazioni di arricchimento della risorsa: le correlazioni paremiologiche tra proverbi affini e la costruzione degli scenari d'uso (triple) che costituiscono DiPRomIt-scenarios. La Sezione 2.3 presenta l'analisi descrittiva del dataset, distinta dall'analisi dei risultati sperimentali del Capitolo 3. La responsabilità primaria sul contenuto è dei contributori WikiQuote, che hanno ceduto i dati con licenza CC BY-SA 4.0; il dataset ne eredita quindi eventuali errori, refusi, problemi di trascrizione o interpretazioni soggettive¹.

2.1 Raccolta dei dati

La raccolta di proverbi italo-romanzi è un'operazione impegnativa, data l'eterogeneità e la variabile affidabilità delle fonti esistenti (libri, raccolte folcloristiche, siti web, post sui social media, *inter alia*) e la scarsità

¹ Coerentemente con la clausola *ShareAlike* della licenza sorgente, anche DiPRomIt viene rilasciato con licenza CC BY-SA 4.0.

di proverbi dialettali che soddisfino contemporaneamente tutti i seguenti requisiti: (i) disponibilità nel dominio pubblico con licenza permissiva; (ii) accompagnamento da una traduzione letterale o dal significato in italiano; (iii) indicazione dell'area geografica o della varietà dialettale romanza di appartenenza; (iv) attestazione scientifica con indicazione esplicita della fonte originale. Per la raccolta ci siamo pertanto affidati all'unica risorsa che soddisfa tutti i criteri sopra elencati: la sezione dei proverbi dialettali dell'edizione italiana di WikiQuote².

Poiché i proverbi di ciascuna varietà sono attestazioni autonome, non traduzioni reciproche, DiPRomIt si configura come un corpus comparabile³ (McEnery & Xiao, 2007): proverbi attestati in varietà diverse, strutturati per consentire confronti utili alla ricerca, soprattutto, ma non limitati, all'asse sincronico e alla dimensione diatopica.

La popolazione del dataset ha avuto inizio nella settimana del 23 marzo 2026 e si è conclusa il 6 aprile 2026 (dump dei dati: `itwikiquote-20260301`). Questa informazione è particolarmente rilevante poiché WikiQuote è un contenuto contributivo dinamico; un ricercatore che accede alle stesse pagine in una data successiva potrebbe trovare voci aggiunte, modificate o rimosse rispetto a quelle presenti durante la raccolta.

Sono state raccolte le voci da tutte le sezioni e sottosezioni disponibili in WikiQuote al momento della raccolta, con l'eccezione della sezione *Pro-*

² https://it.wikiquote.org/wiki/Categoria:Proverbi_nei_dialetti_italiani (consultato il 1° maggio 2026).

³ Si differenzia da un corpus parallelo, in cui gli stessi testi o frasi appaiono tradotti in più lingue. Difatti, i campi Traduzione e Significato non si possono considerare traduzioni dirette nel senso stretto del termine.

verbi toscani, esclusa perché contiene proverbi scritti in italiano standard anziché in dialetto: per questa regione sono stati considerati unicamente i *Proverbi livornesi*, gli unici a soddisfare i criteri di inclusione. Ciascuna sottosezione corrisponde a un'area geografica italiana con una determinata granularità (ad esempio la regione amministrativa nella pagina *Proverbi campani*, oppure il comune nella pagina *Proverbi procidani*, Procida, Napoli, Campania): il titolo geografico di ciascuna sottosezione è stato utilizzato per etichettare i proverbi corrispondenti con informazioni geografiche al massimo livello di granularità disponibile (regione, provincia, comune).

Per vincoli di tempo, le tre sezioni con un numero di voci molto elevato (*Proverbi calabresi*, *Proverbi napoletani*, *Proverbi tarantini*, ciascuna con oltre 500 proverbi) sono state limitate ai primi 50 proverbi disponibili; lavori futuri potranno integrare DiPRomIt con ulteriori voci da queste pagine⁴.

Si è deliberatamente evitato il *web scraping*; l'intera raccolta è stata condotta manualmente, data la variabilità con cui le informazioni sono presentate nelle diverse pagine (ad esempio, traduzioni letterali e significati sono talvolta riportati in ordine invertito o con formattazione non uniforme). La raccolta manuale ha permesso di: (i) aggiungere le informazioni geografiche che un processo automatico non garantirebbe, come gli ID correlati (cfr. Sezione 2.2.1) e alcuni dettagli che nella risorsa primaria erano scritti in nota o a margine; (ii) effettiva documentazione attestata del proverbio; (iii) verifica dell'integrità del testo, delle traduzioni e dei

⁴ La priorità nella compilazione è stata assegnata alle regioni con numerosità contenuta. La soglia minima adottata è di 25 proverbi per regione: le regioni al di sotto di tale soglia, dopo valutazione di fonti integrative, sono state escluse.

significati rispetto alla fonte originale⁵.

Non tutte le fonti candidate hanno soddisfatto i criteri di inclusione elencati in apertura di sezione: le fonti e le regioni scartate sono registrate in un foglio dedicato con indicazione del motivo dell'esclusione; se una sezione WikiQuote riporta un banner di avviso sulla mancanza di fonti, viene esclusa per mancato soddisfacimento del criterio di attestazione documentata (come illustrato più avanti). La Figura 2.1 riassume graficamente le fasi della pipeline appena descritte.

2.1.1 Criteri di inclusione

Riassumendo, un proverbio è incluso nel dataset se soddisfa entrambe le seguenti condizioni obbligatorie:

1. **(C1) Traduzione o significato presenti:** il proverbio deve essere accompagnato almeno da una traduzione letterale o dal suo significato in italiano.
2. **(C2) Attestazione documentata:** deve essere presente almeno una fonte bibliografica o scientifica verificabile, che sia una pubblicazione o un libro che raccolga e documenti proverbi (es. CC BY-SA 4.0).

Questo ha comportato lo scarto di numerose voci con fonti mancanti, non disponibili o inaffidabili (ad esempio, tutti i proverbi della raccolta *Proverbi*

⁵ Questa procedura si è rivelata particolarmente complessa, quindi si è optato per verificare le voci a campione: circa il 10% dei proverbi e delle relative traduzioni e significati in DiPRomIt sono stati controllati dai curatori del dataset. Eventuali refusi o forme ortografiche difformi dall'originale sono stati corretti.

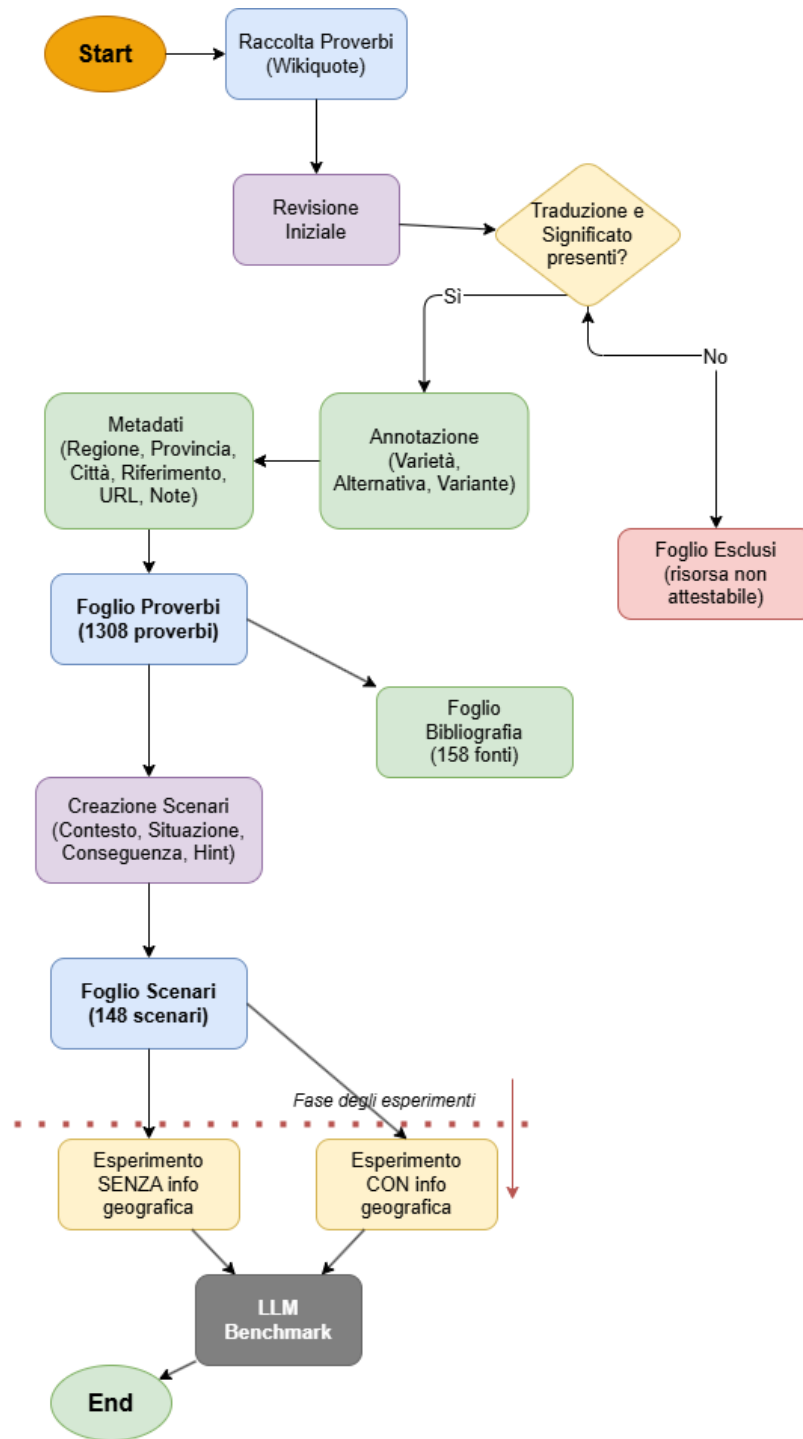


Figura 2.1: Pipeline di raccolta e arricchimento dei dati di DiPRomIt e di costruzione di DiPRomIt-scenarios. I riquadri blu: fogli di lavoro prodotti in ciascuna fase (proverbi, scenari); viola: fasi di elaborazione manuale (revisione, creazione degli scenari); il rombo giallo: punto di decisione basato sui criteri di inclusione; verde: fasi di arricchimento e ristrutturazione dei metadati; rosso: elementi esclusi dalla risorsa. Sotto la linea tratteggiata, le direttive sperimentali (Capitolo 3) sono riportate a titolo di contestualizzazione.

abruzzesi riportanti come fonte il sito «ChietiToday.it») e persino di intere sottosezioni (ad esempio, *Proverbi catanzaresi*, priva di qualsiasi fonte).

Si è scelto di includere anche proverbi provenienti da fonti letterarie, filmiche e gastronomiche. A titolo esemplificativo di questa sottocategoria si presenta il proverbio con ID 523 (Arosio e Maimone, 2013), riportato nel romanzo *Vertigine*:

Proverbio	<i>Acqua ca duna a tanti vadduni a mari non ci n'arriva.</i>
Traduzione	Acqua che va in tanti torrenti non arriva al mare.
Significato	Se si fanno troppe cose non se ne porta a compimento nessuna.

2.1.2 Schema del dataset

Il dataset è strutturato come un file tab-separated in cui ogni riga corrisponde a un proverbio e ogni colonna a una proprietà della voce. A ogni proverbio è assegnato un identificatore univoco. I campi del dataset si dividono in tre gruppi: *contenuto linguistico* (*Proverbio*, *Traduzione*, *Significato*), che cattura il testo del proverbio e il senso annesso; *provenienza geografica* (*Regione*, *Provincia*, *Città*), che ne localizza l'attestazione con la granularità più precisa disponibile; *relazioni e metadati* (*Alternativa*, *Variante*, *Riferimento*, *URL*), che collegano le voci tra loro e alle fonti originali.

La tabella 2.1 riporta lo schema completo con il ruolo di ciascun campo; le sezioni seguenti ne dettagliano i criteri di compilazione.

Tabella 2.1: Schema del dataset. Le righe evidenziate, *Traduzione* e *Significato*, indicano che almeno uno dei due deve essere necessariamente presente. I campi geografici (*Regione*, *Provincia*, *Città*) sono compilati solo quando l'informazione risulta certa.

Campo	Note
ID	Identificativo del proverbio.
Proverbio	Testo del proverbio nella varietà originale.
Traduzione	Significato letterale.
Significato	Interpretazione figurata del proverbio.
Note	Interpretazione personale del contribuente, non riconducibile a una fonte verificabile: spostata qui anziché in <i>Significato</i> .
Regione	Provenienza geografica a livello regionale.
Provincia	Livello provinciale.
Città	Livello comunale.
Alternativa	ID di proverbi sinonimici nella stessa regione.
Variante	ID di proverbi sinonimici in regioni diverse
URL	Link alla fonte online.

Traduzione e Significato

I due campi hanno funzioni distinte: *Traduzione* riporta il significato letterale del proverbio dialettale; *Significato* ne riporta l'interpretazione. In alcuni casi isolati WikiQuote fornisce due significati alternativi separati da «oppure» o «equivalente»: si è scelto di inserirne uno solo nella cella apposita, senza sdoppiare la voce in due ID distinti.

Nei casi in cui ciò che compare sotto al proverbio in WikiQuote sia riconducibile a un'interpretazione personale del contribuente (quindi non a una fonte verificabile), e quindi a un commento, il contenuto viene spostato nella colonna *Note* anziché mantenuto in *Significato*. Se la fonte della traduzione o del significato non è espressamente indicata, si intende coincidente con la fonte del proverbio.

Va segnalato che in alcuni casi i due campi potrebbero tendere alla

sovrapposizione. Non sempre è immediatamente distinguibile se il testo riportato sotto al proverbio nella risorsa originaria costituisca la traduzione o l'interpretazione. A titolo d'esempio si riporta il proverbio veneto con ID 138:

Proverbio	<i>Carta canta, vilan dormi.</i>
Traduzione	<i>Quel che è scritto è quel che vale, mentre il contadino (che normalmente non sapeva scrivere) dorme.</i>
Significato	Non presente.

Come si può notare, qui sono possibili almeno due interpretazioni, entrambe plausibili:

1. il contribuente ha aggiunto un'interpretazione al testo originale, il testo tra parentesi;
2. nella fonte originale non si dà peso alla distinzione tra spiegazioni e interpretazioni concettuali dell'autore, che si mescolano alla pura traduzione del proverbio.

In questo caso abbastanza raro si è optato per tenere l'intero enunciato, incluso l'inciso, che è stato mantenuto nel campo *Traduzione* così come riportato nella tabella precedente, lasciando *Significato* vuoto. Una ricerca a campione ha cercato di minimizzare i potenziali problemi di voci confondibili come queste, tra Traduzione e Significato. Nei due sottoparagrafi seguenti si presentano alcune difficoltà tecniche che si possono riscontrare in lavori simili durante l'organizzazione di dati semi-strutturati; sono documentate nell'Appendice E (§45). Reputiamo utile una rapida presentazione

dei seguenti problemi in modo da rendere le sottosezioni una specie di guida, qualora in futuro si volesse implementare una metodologia analoga.

Campo *Alternativa* Il campo *Alternativa* collega proverbi lessicalmente prossimi ma intra-regionali. Le alternative si indicano come elenco di identificatori numerici separati da «;» (es. «31; 96; 155»), in modo uniforme, indipendentemente dal numero di elementi. Il formato è pensato per consentire il parsing automatico tramite un semplice `split` sul separatore. Se il proverbio A ha come alternativa il proverbio B, nella riga di A compare l’ID di B e nella riga di B compare l’ID di A; i proverbi sono così collegati in modo bidirezionale.

Tabella 2.2: Esempio di voci in relazione di *Alternativa*.

ID	Proverbio	Traduzione
343	<i>Per la gotta è bon nagotta.</i>	<i>n.d.</i>
344	<i>Per la gotta ghe völ el son d’ona campana rotta.</i>	Contro la gotta poco può l’arte.

Campo *Variante* Il campo *Variante* registra proverbi semanticamente equivalenti attestati in regioni o varietà diverse, riflettendo quindi la dimensione diatopica. Si distingue da *Alternativa* (cfr. campo *Alternativa*) per l’ambito extra-regionale della relazione tra ID. La *ratio* sottostante è tenere traccia di variazioni intra ed extra regionali. Anche questo campo segue il formato appena visto nella sezione precedente. Va precisato che questa distinzione si basa sulla granularità regionale: due proverbi di varietà intra-regionali (ad esempio, dialetti provinciali diversi all’interno della stessa regione) rica-

dono comunque nella categoria *Alternativa*, anche quando la loro distanza linguistica effettiva sia significativa.

La tabella 2.3 illustra un caso di collegamento tra proverbi. Tre proverbi attestati rispettivamente in Veneto, Campania e Lombardia condividono lo stesso nucleo semantico: con l'avanzare dell'età dell'uomo (segnalato dai capelli bianchi, dai sessant'anni, dalla barba canuta), essi invitano a privilegiare il vino alle donne. Pur differendo sostanzialmente nella forma, l'immagine evocata e il significato soggiacente restano quindi estremamente simili. Ciascuna voce riporta nel campo *Variante* gli ID delle altre due, realizzando un collegamento analogo a quello descritto precedentemente per il campo *Alternativa* (cfr. tabella 2.2).

Tabella 2.3: Esempio di voci in relazione di *Variante*.

ID	Proverbio	Traduzione	Regione	Provincia	Variante
153	<i>Co 'l cavelo tira el bianchin, lassa la dona e tient al vin.</i>	Quando i capelli cominciano a imbiancare, lascia la donna e aggrappati al vino.	Veneto	<i>n.d.</i>	292; 353
292	<i>Vucin a la ssandina, lassa r fèmmn e piglia lu vin.</i>	Con l'avvicinarsi dei sessant'anni, lascia le donne e prendi il vino.	Campania	Salerno	153; 353
353	<i>Quando la barba la trà al bianchi, lassa la dònna e ciapa 'l vi.</i>	Quando la barba comincia a incanutire, lascia la donna e prendi il vino.	Lombardia	Bergamo	153; 292

In breve il criterio di disambiguazione *Variante/Alternativa* adottato è il significato: se due proverbi hanno significato pressoché identico e provengono da regioni diverse, vengono collegati come *Varianti*; se risultano lessicalmente simili ma della stessa regione, come *Alternative*.

Campi Provenienza geografica: Regione, Provincia e Città

Qui sono stati raccolti gli attributi geografici di provenienza ordinabili per regione, area, comune e città, così come sono stati riportati nella pagina WikiQuote originale. La granularità disponibile varia da voce a voce: non tutte le fonti riportano lo stesso livello di dettaglio. Sono stati inseriti solo gli attributi effettivamente presenti al momento della raccolta.

Durante l'acquisizione dei dati si sono riscontrate aree storiche o geografiche non corrispondenti a comuni; in questi casi si è scelto di usare il nome della zona senza cercare una provincia o regione amministrativa equivalente, poiché ricondurle a un'unità amministrativa specifica sarebbe risultato in una forzatura. Tutti i territori di questo tipo (Gallura, Monferrato, Irpinia, Brianza, Valtellina, Romagna, Val Camonica, Basilicata meridionale, Salento) sono preceduti da un asterisco (*) per facilitarne l'identificazione in fase di analisi⁶. L'asterisco quindi segnala la peculiarità di queste zone al lettore del dataset. Si sono inoltre riscontrati, in alcuni casi, problemi di visualizzazione legati alla codifica ortografica⁷ di alcune varietà dialettali.

In rari casi (come illustrato nella Tabella 2.4), può accadere che il corrispondente italiano di un proverbio, quando si tratta di una variante ben nota e diffusa, venga riportato, come da fonti WikiQuote, direttamente nel campo *Traduzione*. In tali circostanze il corrispondente italiano non viene

⁶ Comprensori di questo tipo sono riconducibili, con dovuta cautela al singolo caso, al concetto di *transition area*, cioè zone di transizione la cui varietà linguistica risente dell'influsso delle aree confinanti (Kurath, 1949).

⁷ Alcune convenzioni ortografiche utilizzate per alcune varietà dialettali italiane possono presentare caratteri non visualizzabili correttamente in tutti gli ambienti di raccolta dati; tali casi vanno ispezionati a priori e segnalati.

spostato nel campo *Note*, come da prassi.

Tabella 2.4: Esempio di voce con corrispondenza italiana nota.

ID	Proverbio	Corrispondente/Significato	Regione
435	<i>Gialine viere fās bon brūt</i>	Gallina vecchia fa buon brodo	Friuli-Venezia Giulia

Campo Fonti bibliografiche

Ogni fonte citata nel dataset ha una corrispondenza nel foglio della bibliografia. Nel caso in cui ci fossero due opere dello stesso autore e dello stesso anno, si disambigua in linea con le norme editoriali accademiche più diffuse Anno-Lettera, ad esempio Pasqualigo (2026a) e Pasqualigo (2026b). Molte delle fonti bibliografiche citate su WikiQuote sono testi antichi, collocabili prevalentemente tra la fine dell'Ottocento e l'inizio del Novecento. Una parte di esse risulta fuori catalogo, non digitalizzata o di difficile accesso tramite i sistemi bibliotecari disponibili. Questa limitazione ha reso impossibile una verifica sistematica della corrispondenza tra il testo del proverbio riportato su WikiQuote e la fonte originale; alcune voci sono state controllate a campione tramite Google Libri o link disponibili al momento della consultazione.

2.2 Arricchimento della risorsa

Oltre alle informazioni geografiche, DiPRomIt è arricchito manualmente con dati aggiuntivi sulla variazione paremiologica all'interno e tra le diverse

aree geografiche. Come anticipato, per ciascun proverbio e ove possibile, sono stati annotati due tipi di relazione tra voci: (i) *alternative*: altri proverbi in DiPRomIt che condividono il medesimo significato del proverbio di riferimento, attestati nella stessa area geografica (cfr. campo *Alternativa*); e (ii) *varianti*: proverbi con la stessa traduzione letterale o significato del proverbio di riferimento, ma espressi in un'altra varietà romanza parlata in un'area geografica diversa all'interno dell'Italia (cfr. campo *Variante*).

Questo tipo di annotazione prosegue, su base computazionale, l'interesse per la distribuzione geografica del proverbio che caratterizza la *geoparemiologia* di Franceschi (cfr. Sezione 1.1), qui rappresentata in particolare dall'Atlante Paremiologico Italiano (Franceschi, 1994; Franceschi et al., 2000). Queste informazioni possono infatti rendersi utili, ad esempio, per la costruzione di atlanti paremiologici computazionali e per lo studio della variazione diatopica dei proverbi.

L'intero processo di raccolta è stato validato e verificato da due curatori madrelingua italiani con formazione in linguistica, informatica e variazione linguistica nel NLP, attraverso un ciclo iterativo di discussione asincrona⁸. Un esempio di istanze di dati presenti in DiPRomIt è riportato nella

Tabella 2.5.

⁸ Poiché la raccolta e l'arricchimento dei dati hanno comportato la curazione e l'integrazione di informazioni oggettivamente verificabili, seguite da un processo iterativo di consenso tra i due curatori, il calcolo dell'accordo inter-annotatore è stato ritenuto non informativo ai fini della valutazione della qualità dei dati (Artstein & Poesio, 2008).

⁹ Interessante è notare l'esistenza di una versione toscana di questo proverbio, «Al primo colpo non cade l'albero», che compare, nella forma «per lo primo colpo non cade la quercia», nel *Decameron* di Boccaccio (Giornata VII, Novella IX) (Boccaccio, 1980).

Tabella 2.5: Due istanze di proverbi in DiPRomIt con i relativi metadati.

	ID: 389	ID: 829
Proverbio	<i>Al prem còulp an casca l'alber.</i> ⁹	<i>Tumazzu duru nun pigghia sala-mora.</i>
Traduzione	Al primo colpo non cade l'albero.	Il formaggio duro non prende più sale.
Significato	Bisogna tentare più di una volta per ottenere qualcosa.	Una persona invecchiata non è più in grado di cambiare.
Regione	Emilia-Romagna	Sicilia
Provincia/Area	Bologna	Trapani
Comune	—	Erice
ID Alternativa/e	—	828; 830
ID Variante/i	500; 840	513
Fonte	Berti (1874)	Pitrè (1880)

2.2.1 Correlazioni paremiologiche

La sezione che segue presenta varianti e alternative accomunate da una struttura semantica di base. Per la differenza tra le due etichette vedere la sottosezione dedicata in questo capitolo.

Gruppo di varianti VINO → LATTE

Si riportano come casi studio solo gruppi di correlazioni; l'elenco completo delle correlazioni paremiologiche annotate in DiPRomIt è disponibile nella risorsa scaricabile su richiesta (Errico, 2026). Il gruppo, catalogato a partire dall'ID 372, è costituito da sei proverbi che condividono il frame¹⁰ concettuale VINO → LATTE.

In questo esempio, il vino svolge per l'anziano la funzione che il latte materno svolge per il neonato, e in sei regioni diverse con sei diverse rappresentazioni lessicali del concetto. I termini impiegati per designare il

¹⁰ La nozione di «frame» qui impiegata richiama quella di frame semantico in FrameNet (Fillmore & Baker, 2012).

referente (*tetta, tatta, pupa, tettin, latt*) riflettono due gradazioni di registro e trasparenza metaforica. In particolare, si noti la variante friulana (ID 434), che si distingue dalle altre per un maggior grado di esplicitezza. Nelle altre varianti, invece, l'analogia vino/latte rimane implicita.

Tabella 2.6: Gruppo di varianti VINO → LATTE. 6 proverbi

ID	Proverbio	Traduzione	Regione	Tipo
372	<i>El vin l'è la tetta di vecc.</i>	Il vino è la mammella dei vecchi.	Lombardia	Variante
390	<i>Al vén l'è la tatta di vîc.</i>	Il vino è il biberon dei vecchi.	Emilia-Romagna	Variante
434	<i>El vin al è latt da puars vecios.</i>	Il vino è il latte dei poveri vecchi.	Friuli-V.G.	Variante
492	<i>Er vino è la zinna de li vecchi.</i>	Il vino è la mammella dei vecchi.	Lazio	Variante
508	<i>El vin a l'è la pupa dii vei.</i>	Il vino è la mammella dei vecchi.	Piemonte	Variante
945	<i>U vin l'è u tettin di vègi.</i>	Il vino è la mammella dei vecchi.	Liguria	Variante

Gruppo di varianti LA FAME FA USCIRE IL LUPO

Viene riportato di seguito un secondo esempio di analisi preliminare-esplorativa di relazione inter-regionale: il gruppo 1122. Rispetto al gruppo precedente VINO → LATTE, questo caso è simile ma leggermente diverso: mentre per VINO → LATTE i domini tendevano a un lessico concreto (vino, latte), qui i sostantivi (fame, lupo) richiamano direttamente un frame più astratto, di natura semantica: NECESSITÀ VINCE PRUDENZA.

Tabella 2.7: Gruppo di varianti LA FAME FA USCIRE IL LUPO. 2 proverbi

ID	Proverbio	Traduzione	Regione	Tipo
1122	<i>'A fami caccia 'u lupu d''u boscu.</i>	La fame caccia il lupo dal bosco.	Calabria	Variante
1187	<i>'A famma fa asci 'o lupu, da 'o bosco...</i>	La fame fa uscire il lupo dal bosco...	Campania	Variante

2.2.2 Creazione degli scenari

Per valutare la sensibilità culturale dei LLM è stato costruito manualmente DiPRomIt-scenarios, un dataset di scenari a scelta multipla. Per ciascun proverbio selezionato da DiPRomIt è stato redatto uno scenario narrativo, pensato per elicitarlo quel proverbio in contesto, e associato a quattro opzioni di risposta. Ogni scenario con le sue quattro opzioni costituisce un'istanza di DiPRomIt-scenarios e, per costruzione, è pensato per testare la *sensibilità culturale contestuale* dei modelli piuttosto che la sola memorizzazione (ad esempio tramite task di completamento del proverbio).

Gli identificatori degli scenari in DiPRomIt-scenarios costituiscono ID indipendenti rispetto agli identificatori dei proverbi in DiPRomIt: un ID scenario e un ID proverbio con lo stesso valore numerico non fanno riferimento alla stessa voce. Ogni scenario esiste in due versioni parallele: una che include l'informazione sulla provenienza geografica del proverbio (*with-geo*) e una che non la include (*without-geo*). I criteri sono discussi più avanti in questa sezione. Un esempio di istanza di DiPRomIt-scenarios è riportato nella Tabella 2.8.

Tabella 2.8: Esempio di istanza di DiPRomIt-scenarios (versione *with-geo*: l'informazione geografica è sottolineata).

Scenario ID: 147 (basato sul proverbio ID: 679)	
Una bambina è con il papà e la nonna al mercato <u>in Sicilia</u> . La bambina nota della cioccolata in un banco del mercato ed insiste per ottenerla, dopodiché dice al papà: «Sei il papà più bravo del mondo!» Il papà si fa lusingare dalla bambina e le compra il cioccolato tanto agognato. Che proverbio userebbe la nonna in questa situazione?	
A) Bona parola è menzu pagamentu.	✗ W_{same} (Sicilia)
B) La Carni di l'omu si mancia cu li meli.	✓ C (Sicilia)
C) Scinnarö nêvusö, annö früttuosö.	✗ $W_{distant}$ (Campania, Avellino)
D) A 'a casa du 'mpisu no 'mpendiri ogghialòra.	✗ W_{close} (Calabria, Reggio Calabria)

Per la creazione degli scenari sono state definite delle linee guida ancor prima di procedere alla redazione dei singoli scenari. Gli scenari sono composti da tre parti: il Contesto (personaggi/luogo), la Situazione/azione e la Conseguenza/risultato, che chiameremo *triple*. Infine è stato stipulato un suggerimento (*hint*) standardizzato per tutte le triple, della forma: *Che proverbio userebbe [agente/persona] in questa situazione?*, dove [agente/persona] è solitamente introdotta nella tripla come uno dei personaggi, di solito il più anziano.

Per la creazione delle *triple* si è preferito dare priorità ai proverbi che avessero già il significato associato o per i quali si conoscesse già l'utilizzo contestuale. Ove possibile si sono evitati proverbi dal significato troppo letterale: essendo probabilmente più in uso in italiano, la loro rappresentazione nella conoscenza dei LLM è quasi certamente superiore o deducibile dai dati di *training* rispetto a quei proverbi per cui la memorizzazione non basta; l'obiettivo invece rimane quello di spingere i modelli a «ragionare» piuttosto che ricorrere alla sola memorizzazione.

Le *triple* devono essere sufficientemente neutre, brevi, generiche e prive di dettagli superflui. Questo in modo da permettere, congiuntamente, di inferire da parte del modello il proverbio più pertinente. Il proverbio, o parte di esso, non deve assolutamente essere menzionato direttamente: ciò permetterebbe al modello di riprodurlo come semplice continuazione di una sequenza testuale già vista in fase di addestramento, vanificando l'obiettivo della ricerca. Siccome creare degli scenari che elicitino un proverbio non

è sempre immediato, qualora la formulazione risultasse particolarmente difficoltosa, è stata ammessa, in via del tutto eccezionale, e con uso il più limitato possibile, l'opportunità di includere una parola del proverbio.

Il testo di ciascuna parte della tripla include gli elementi *variabili*, ovvero i personaggi e i luoghi, rispettivamente tra parentesi quadre e parentesi graffe, al fine di consentire l'estensione programmatica degli scenari ad altri referenti in modo completamente automatico generando elementi casuali con quelli inseriti manualmente in questa fase iniziale. Questa era inizialmente la finalità principale; l'esperienza pratica di cura del dataset ha tuttavia dimostrato che tale operazione, considerata in un primo momento immediata, si è rivelata particolarmente ardua, un limite noto della generazione basata su template rispetto all'accordo di genere grammaticale (Aggarwal et al., 2026), ed è stata pertanto abbandonata. Per completezza, si è scelto di continuare a utilizzare la convenzione inizialmente adottata, anche perché permette di riconoscere gli agenti, i pazienti e i luoghi a colpo d'occhio.

La ricerca si posiziona nella più ampia branca della *awareness*, di conseguenza si è pensato che per la creazione degli scenari si dovesse fare particolare attenzione¹¹, come anticipato, di non fare ricorso a stereotipi, anche impliciti, privilegiando formulazioni neutre a concetti pregiudizievoli, polarizzati, stereotipati (cfr. Mieder, 2004, p. 28) e, in generale, distorti che

¹¹ È stata inoltre verificata la correttezza della punteggiatura di ciascuno scenario, assicurandosi che ogni frase si concludesse con un punto: la segmentazione automatica delle frasi (*sentence boundary detection*) si basa infatti principalmente su segnali di punteggiatura e sullo spazio bianco che segue il punto, ed errori di questo tipo si propagherebbero alle fasi successive della pipeline di elaborazione (Read et al., 2012).

avrebbero potuto influire sull'outcome del modello (es. «*[un uomo] al bar di paese [in Trentino-Alto Adige]*» invece di «*un ubriacone della regione x*»).

Contesto (personaggi/luogo)

Per la creazione del Contesto si è scelto di non utilizzare nomi propri per i personaggi, bensì ruoli generici legati alla sfera familiare e amicale (es. *il padre, la nonna, l'amico*). Il formato adottato segue il modello:

[Una ragazza] è con [suo nonno] in un campetto da pallacanestro
{in Emilia-Romagna}.

Situazione/azione

La Situazione/azione descrive il comportamento dei personaggi: la situazione in cui si trovano (presente o passata) e l'azione, fondata sul significato del proverbio, che determina una conseguenza o un risultato. In questa parte della tripla è necessario indicare esplicitamente il personaggio coinvolto, al fine di evitare ambiguità tra i referenti e non penalizzare il modello in modo non pertinente.

Conseguenza/risultato

La Conseguenza/risultato specifica l'effetto, positivo o negativo, che rende naturale il ricorso al proverbio da parte dell'altro personaggio. Anche in

questa sezione il personaggio deve essere indicato esplicitamente per evitare ambiguità.

Hint

Il suggerimento (*hint*) specifica sempre il personaggio che pronuncia il proverbio.

Costruzione delle opzioni

Per ogni scenario sono state selezionate quattro opzioni proverbiali:

- **C** (*Corretta*): il proverbio che lo scenario elicitamente direttamente; costituisce la risposta corretta che un LLM dovrebbe prevedere.
- **W_{same}** (*Errata, stessa regione*): il modello ha scelto un proverbio della stessa regione amministrativa di C;
- **W_{close}** (*Errata regione vicina*): il proverbio scelto non è di C, ma di una regione adiacente a C¹²;
- **W_{distant}** (*Errata regione lontana*): il modello ha predetto un proverbio da una regione che non è né quella di C né fa parte delle regioni vicine impiegate per il campionamento di W_{close}.

Nella creazione delle opzioni errate del benchmark si escludono esplicitamente tutti i proverbi classificati come varianti o alternative di C (cfr. campo

¹² Ad esempio Lombardia al posto di Piemonte, in altri casi, come Puglia e Sicilia, tra gli altri (cfr. Tabella 2 in Appendice), il discriminante non è la barriera fisica-geografica: pur non essendo geograficamente contigua alla Sicilia, la Puglia è inclusa nell'insieme W_{close} per i proverbi siciliani, poiché i dialetti romanzi del Salento appartengono allo stesso gruppo linguistico del dialetto siciliano.

Variante), per evitare falsi negativi che porterebbero a una sottostima delle performance dei modelli.

Le opzioni di ciascuno scenario sono state mescolate prima della valutazione per evitare il *position bias*, fenomeno per cui i LLM tendono a favorire determinate posizioni nelle domande a scelta multipla (Pezeshkpour & Hruschka, 2024).

L'intero processo di costruzione e validazione degli scenari è stato condotto dagli stessi due curatori che hanno curato DiPRomIt, entrambi madrelingua italiani con formazione in linguistica, informatica e variazione linguistica nel NLP. Complessivamente sono stati creati 294 scenari paralleli: 147 nella versione con informazione geografica (*with-geo*) e 147 nella versione senza (*without-geo*). I criteri per la selezione dei proverbi su cui costruire le istanze sono stati: (i) il proverbio deve essere accompagnato dal significato o quest'ultimo deve essere noto ai curatori; (ii) il proverbio non deve essere offensivo nei confronti di alcun gruppo; (iii) ove possibile, tutte le regioni devono essere rappresentate in DiPRomIt-scenarios, con un limite di 10 scenari per regione¹³.

2.3 Analisi dei dati

Le statistiche riassuntive di DiPRomIt sono riportate nella Tabella 2.9. Tutte le regioni amministrative italiane sono rappresentate in DiPRomIt, con

¹³ La distribuzione regionale degli scenari è la seguente: Abruzzo (7), Basilicata (10), Calabria (10), Campania (10), Emilia-Romagna (10), Friuli-Venezia Giulia (11), Lazio (10), Liguria (10), Lombardia (10), Marche (2), Molise (4), Piemonte (6), Puglia (10), Sardegna (10), Sicilia (10), Toscana (7), Veneto (10).

l’eccezione di Umbria e Valle d’Aosta, per le quali WikiQuote non rende disponibili proverbi dialettali con fonti verificabili. Complessivamente, DiPRomIt conta **1.310 proverbi**, di cui 1.298 (99,1%) sono accompagnati da una traduzione letterale e 414 (31,6%) dal significato. Il numero totale di token è 14.889, con un vocabolario di 4.585 tipi, cioè forme distinte; la lunghezza media dei proverbi è di 11,4 token¹⁴. Il numero complessivo di fonti bibliografiche distinte è 157.

Tabella 2.9: Statistiche riassuntive di DiPRomIt per regione. *P*: istanze di proverbi (#: numero; %: percentuale sul totale); *T_{all}*: token totali; *T_{uniq}*: token unici; *T/P*: media token per proverbio; *L*: traduzione letterale (# e % dei proverbi con traduzione); *M*: significato (# e % dei proverbi con significato); *S*: numero di fonti.

Regione	Istanze		Token			Traduzione / Significato				Fonti <i>S</i>
	<i>P</i> (#)	<i>P</i> (%)	<i>T_{all}</i>	<i>T_{uniq}</i>	<i>T/P</i>	<i>L</i> (#)	<i>L</i> (%)	<i>M</i> (#)	<i>M</i> (%)	
Abruzzo	26	2,0%	320	158	12,3	26	100,0%	0	0,0%	4
Basilicata	118	9,0%	1.656	619	14,0	117	99,2%	15	12,8%	17
Calabria	82	6,3%	893	399	10,9	82	100,0%	34	41,5%	20
Campania	120	9,2%	1.578	667	13,2	120	100,0%	58	48,3%	28
Emilia-Romagna	18	1,4%	209	126	11,6	18	100,0%	5	27,8%	8
Friuli-Venezia G.	81	6,2%	853	340	10,5	81	100,0%	9	11,1%	4
Lazio	16	1,2%	178	103	11,1	16	100,0%	2	12,5%	7
Liguria	84	6,4%	851	369	10,1	83	98,8%	44	52,4%	11
Lombardia	65	5,0%	736	342	11,3	63	96,9%	26	40,0%	22
Marche	3	0,2%	38	31	12,7	3	100,0%	0	0,0%	2
Molise	10	0,8%	127	78	12,7	10	100,0%	3	30,0%	2
Piemonte	13	1,0%	167	100	12,8	13	100,0%	3	23,1%	6
Puglia	100	7,6%	1.169	478	11,7	100	100,0%	53	53,0%	13
Sardegna	83	6,3%	771	343	9,3	83	100,0%	33	39,8%	8
Sicilia	353	26,9%	3.837	1.079	10,9	353	100,0%	113	32,0%	27
Toscana	9	0,7%	100	61	11,1	1	11,1%	8	88,9%	1
Trentino-A.A.	3	0,2%	30	23	10,0	3	100,0%	0	0,0%	1
Veneto	126	9,6%	1.376	481	10,9	126	100,0%	8	6,3%	3
Totale	1.310	100,0%	14.889	4.585	11,4	1.298	99,1%	414	31,6%	157

La distribuzione delle istanze varia sensibilmente tra le regioni. Ciò riflette la diversa disponibilità di dati affidabili su WikiQuote. La Sicilia è la regione più rappresentata (353 proverbi; 26,9%), seguita da Veneto (126;

¹⁴ Per la tokenizzazione è stato impiegato il modello `it_core_news_sm` di spaCy v3.8 (<https://spacy.io/>).

9,6%) e Campania (120; 9,2%). Le regioni meno rappresentate (meno di 10 proverbi ciascuna) sono Marche, Toscana e Trentino-Alto Adige. In media, i proverbi più lunghi in termini di token appartengono alla Basilicata (14,0 token per proverbio), mentre quelli più brevi sono della Sardegna (9,3 token). Quasi tutte le regioni presentano una traduzione letterale per la quasi totalità dei proverbi ($\geq 96,9\%$), con l'eccezione della Toscana, dove la prossimità dei dialetti locali all'italiano standard, in seguito al riconoscimento del volgare fiorentino come norma di riferimento nel Cinquecento (cfr. Berruto & Cerruti, 2015), rende spesso la traduzione ridondante. La copertura del campo *Significato* varia invece in modo più marcato tra le regioni, passando dallo 0% (Abruzzo, Marche, Trentino-Alto Adige) all'88,9% (Toscana).

2.3.1 Distribuzione geografica e sottoinsieme di valutazione

La distribuzione dei 1.310 proverbi raccolti in DiPRomIt per regione, già riportata nella Tabella 2.9, non riflette una scelta editoriale bensì una disponibilità, disomogenea, dei dati su WikiQuote. Nella Figura 2.2 si mostra il sottoinsieme di proverbi che è stato utilizzato per costruire gli scenari. L'obiettivo era raggiungere almeno 10 scenari per regione, a garanzia di rappresentatività.

Gli scenari sono stati costruiti in modo da normalizzare parzialmente la distribuzione regionale rispetto al corpus completo (DiPRomIt): le regioni sovrarappresentate in DiPRomIt (soprattutto la Sicilia) sono state volutamente limitate a 10 scenari ciascuna, così da rendere la distribuzione

del sottoinsieme di valutazione (DiPRomIt-scenarios) più uniforme tra le regioni.

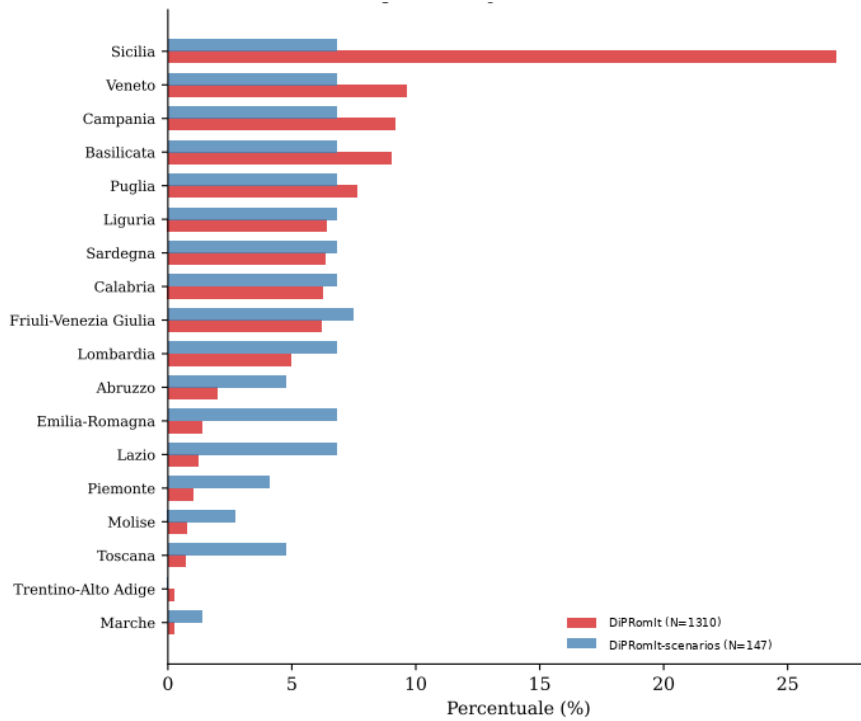


Figura 2.2: Grafico a barre orizzontali affiancate che rappresenta la distribuzione delle istanze in ogni regione presente in DiPRomIt-scenarios (N = 147 scenari), in blu, contrapposte a quelle dell'intero DiPRomIt, in rosso.

Capitolo 3

Esperimenti

Il presente capitolo descrive l'impostazione sperimentale adottata per valutare la sensibilità culturale (*cultural sensitivity*) dei LLM sul benchmark DiPRomIt-scenarios. Il capitolo si articola in tre sezioni. La Sezione 3.1 illustra l'impostazione sperimentale: il formato del task, i modelli impiegati, i prompt, le condizioni geografiche e le metriche di valutazione. La Sezione 3.2 presenta i risultati complessivi su DiPRomIt-scenarios e ne discute i principali pattern. La Sezione 3.3 approfondisce l'analisi degli errori.

3.1 Impostazione degli esperimenti

Riprodurre con precisione l'impostazione sperimentale è condizione necessaria per la validità dei risultati discussi nel resto del capitolo: le sottosezioni seguenti ne dettagliano gli elementi costitutivi.

3.1.1 Il task

Il task si presenta relativamente semplice: al modello viene formulata una domanda a scelta multipla. Per ciascuno dei 147 scenari di DiPRomIt-

scenarios, il modello deve selezionare, tra quattro alternative etichettate con A, B, C, D, il proverbio che meglio si adatta al contesto descritto nella tripla. Le quattro opzioni comprendono il proverbio corretto (C) e i tre distrattori (W_{same} , W_{close} , W_{distant}) descritti nel Capitolo 2. Tutti gli esperimenti sono condotti in regime *zero-shot*¹.

3.1.2 I modelli

Per gli esperimenti si impiegano 10 LLM *instruction-tuned*² *open-weight*³, liberamente disponibili, selezionati per favorire la riproducibilità, nelle loro versioni piccole (7–9B parameters, cioè 7-9 miliardi di parametri) e grandi (27–72B parameters): Llama-3.1-8B e Llama-3.1-70B (Grattafiori et al., 2024), Qwen-2.5-7B e Qwen-2.5-72B (Qwen Team, 2025a), Mistral-0.3-7B (Jiang et al., 2023) e Mixtral-0.1-8x7B (Jiang et al., 2024), Gemma-2-9B e Gemma-2-27B (Gemma Team et al., 2024). In aggiunta, considerato il ruolo centrale dell’italiano come varietà di riferimento nel

¹ Con *regime zero-shot* si intende una modalità di valutazione in cui il modello riceve unicamente l’istruzione del task e l’istanza da risolvere, senza alcun esempio illustrativo incluso nel prompt del comportamento atteso in fase di generazione dell’output: a differenza dell’addestramento supervisionato ad esempio, nello *zero-shot* non vi è qui alcun aggiornamento dei parametri né alcuna esposizione preventiva a coppie input-output specifiche del task. Il modello può fare leva solo a partire dalla conoscenza acquisita in fase di pre-addestramento. Il regime *zero-shot* si contrappone al *few-shot*, in cui alcuni esempi risolti sono inclusi nel prompt/input (Thapa et al., 2025).

² Modelli come GPT-3.5-turbo, PaLM2 chat-bison-001 e Falcon-7B-Instruct (Atreja et al., 2024), sottoposti a (*fine-tuning*) e progettati per insegnargli a seguire istruzioni espresse in linguaggio naturale (Pawar et al., 2025b)

³ Con *open-weight* si intendono i pesi del modello. Essi sono pubblicamente accessibili — di fatto, *open* — per il download e l’utilizzo locale/privato, contrapponendosi ai modelli proprietari accessibili solo via API. Le condizioni di licenza variano in base ai modelli impiegati: Qwen-2.5 e Mistral/Mixtral adottano la licenza Apache 2.0; Gemma-2 è soggetto ai *Gemma Terms of Use* di Google; Llama-3.1, Llamantino-3-ANITA-8B e Minerva-1.0-7B sono soggetti alla *Meta Llama 3 Community License*. Alcuni di essi richiedono autorizzazione preventiva, semplicemente usufruibile con una richiesta tramite il portale del modello.

L’accesso ai modelli con licenza ristretta è stato ottenuto tramite autenticazione accordata previa richiesta email dai singoli distributori dei modelli su Hugging Face.

panorama romanzo italiano, si sono anche testati due modelli con maggiore rappresentazione dell'italiano nei dati di pre-addestramento: Llamantino-3-ANITA-8B (Polignano et al., 2024) e Minerva-1.0-7B (Orlando et al., 2024), disponibili solo nella fascia parametrica ridotta.

I modelli sono stati scaricati dall'hub di Hugging Face e stanziati tramite la libreria `transformers` (Wolf et al., 2020). I modelli di dimensioni ridotte sono caricati con *quantizzazione* a 8 bit tramite `bitsandbytes`; i modelli di dimensioni maggiori sono quantizzati a 4 bit. La quantizzazione è la tecnica con cui i pesi del modello vengono rappresentati con un numero minore di bit rispetto al loro formato originale (che tipicamente corrisponde a 16 o 32 bit). Ciò permette di ridurre il carico sulla GPU. Il costo è una perdita di precisione numerica, e di conseguenza una perdita di performance, anche se generalmente trascurabile in entrambi i casi. Quanto più bassa è la quantizzazione in bit, tanto più alte sono le prestazioni della GPU, anche se ciò fa scendere leggermente quelle del modello. Invece, `bitsandbytes` è la libreria software che implementa questa riduzione.

La MoE rappresenta un'architettura in cui il modello è composto da varie sotto-reti specializzate («experts» — esperti), di cui solo un sottoinsieme viene attivato per ciascun input, ove richiesto. Seppur risultando più efficiente in fase di calcolo, data l'attivazione singola o asincrona di singole parti, l'architettura richiede comunque di mantenere in memoria *tutti* gli esperti, anche quelli non attivati in un dato passaggio: da qui il maggiore

fabbisogno di memoria che giustifica una quantizzazione più bassa⁴. Gli esperimenti sono stati eseguiti in un ambiente containerizzato Docker su GPU NVIDIA Tesla V100-SXM2-32GB (modelli di dimensioni ridotte) e NVIDIA A40-48GB (per i modelli di dimensioni maggiori).

I valori degli iperparametri⁵ (Mensa et al., 2025) sono quelli predefiniti di ciascun modello; la generazione è vincolata a un singolo token di output (A, B, C o D), con l’eccezione di Gemma-2-9B e Minerva-1.0-7B, per i quali si è impostato `max_new_tokens=2`⁶ per gestire un carattere di newline aggiuntivo che questi modelli tendono a preporre alla risposta. Gli identificatori ufficiali, il numero di parametri e la provenienza di ciascun modello sono riportati in Tabella 3.1.

Tabella 3.1: Modelli utilizzati negli esperimenti, con rispettivi identificatori Hugging Face, numero di parametri e origine.

Nome breve	Identificatore Hugging Face	Param.	Origine
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct	8B	Meta
Gemma-2-9B	google/gemma-2-9b-it	9B	Google
Qwen-2.5-7B	Qwen/Qwen2.5-7B-Instruct	7B	Alibaba
Mistral-0.3-7B	mistralai/Mistral-7B-Instruct-v0.3	7B	Mistral AI
Llamantino-3-ANITA-8B	swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA	8B	SWAP-Uniba
Minerva-1.0-7B	sapienzanlp/Minerva-7B-instruct-v1.0	7B	SapienzaNLP
Llama-3.1-70B	meta-llama/Llama-3.1-70B-Instruct	70B	Meta
Gemma-2-27B	google/gemma-2-27b-it	27B	Google
Qwen-2.5-72B	Qwen/Qwen2.5-72B-Instruct	72B	Alibaba
Mixtral-0.1-8x7B	mistralai/Mixtral-8x7B-Instruct-v0.1	8×7B	Mistral AI

Si è inoltre scelto di non includere modelli proprietari accessibili solo tramite

⁴ Per approfondire questo aspetto tecnico, cfr. Dettmers e Zettlemoyer (2022) e Dettmers et al. (2022).

⁵ Gli iperparametri sono configurazioni esterne al modello, preimpostate e non apprese in fase di allenamento del modello: si distinguono così dai *parametri*, i pesi del modello appresi durante l’addestramento e già introdotti precedentemente (cfr. paragrafo «Modelli», il numero che dà e segue il nome di ciascun modello, ad es. 7B, 70B). Si possono a loro volta distinguere due tipi di iperparametri: quelli del *training setup*, fissati in preaddestramento per definirne il comportamento, come il *learning rate* (es. 0,00002) (Ghosh & Srivastava, 2022); e quelli dell’*inference/generation setup* (*ibidem*), fissati prima dell’esecuzione per controllare l’output, come il seed casuale (es. 42), impostato in questa fase non essendovi un vero e proprio addestramento.

⁶ Questa scelta riprende quanto osservato nei test preliminari: senza il margine aggiuntivo, la risposta effettiva veniva troncata e l’output risultava sistematicamente non valido.

API (ad es. GPT-4, Claude, Gemini): l’invio degli scenari di DiPRomIt-scenarios a servizi commerciali chiusi ne comporterebbe l’esposizione a infrastrutture esterne le cui politiche di conservazione e riutilizzo dei dati in ingresso non sono verificabili né garantite, con il rischio di una futura contaminazione del benchmark qualora gli scenari venissero incorporati nei dati di addestramento di versioni successive degli stessi modelli.

3.1.3 I prompt

È noto che i LLM sono altamente sensibili alla formulazione dell’input, specialmente in contesti zero-shot (Liu et al., 2024). Per questo motivo, anziché affidarsi a un singolo prompt, sono state progettate tre varianti dell’istruzione base, che variano lungo due assi: *verbosità dell’istruzione* e *framing del task*. Le parole sono state selezionate con cura, poiché differenze anche minime nella formulazione possono condizionare i risultati (Mizrahi et al., 2024). Tutti i prompt sono formulati in italiano, poiché lingua-tetto per la maggior parte dei dialetti italo-romanzi parlati in Italia:

Template prompt (it)

\$ISTRUZIONE_BASE

\$SCENARIO

\$OPZIONI

Non includere commenti o caratteri speciali alla risposta.

Le risposte possibili sono solo A, B, C, D.

Risposta:

Un'istanza di DiPRomIt-scenarios è composta dalla tripla (Sezione 2.2.2) e dalle relative quattro opzioni di risposta. Nel template, \$SCENARIO corrisponde al testo della tripla, \$OPZIONI è l'elenco delle quattro opzioni disponibili etichettate A–D, costruite come descritto nella sezione *Costruzione delle opzioni* del Capitolo 2, e \$ISTRUZIONE_BASE è una delle seguenti varianti:

1. «*Dato il seguente contesto comunicativo, scegli il proverbio dialettale più appropriato tra le opzioni proposte.*» — l'istruzione più diretta, prototipo delle altre due. Include il qualificatore *dialettale* e inquadra lo scenario come *contesto comunicativo*.
2. «*Scegli il proverbio più appropriato.*» — versione ridotta all'essenziale, priva di framing esplicito e del qualificatore *dialettale*. È progettata per rispondere alla domanda: in assenza di indicatori espliciti del task, in che misura cambiano i risultati?
3. «*Scegli il proverbio dialettale più appropriato tra le opzioni proposte data la seguente scena comunicativa.*» — variante che inverte l'ordine sintattico del Prompt 1, mantiene il qualificatore e sostituisce *contesto comunicativo* con *scena comunicativa*⁷.

⁷ La scelta di alternare questi due termini *contesto comunicativo* e *scena comunicativa* nella valutazione degli output non è casuale. Entrambi questi termini sono riconducibili alla nozione di *speech event* (Hymes, 1974). I due sintagmi qui utilizzati però tendono ad attivare frame concettuali diversi nell'essere umano (il primo più ipotetico-situazionale, il secondo più pragmatico-narrativo). La loro inclusione come variabile di prompt consente di verificare se questa differenza incida o meno sulle risposte del modello in modo quantificabile.

Ogni prompt termina con il medesimo vincolo di output: «*Non includere commenti o caratteri speciali alla risposta. Le risposte possibili sono solo A, B, C, D.*», per elicitare una risposta breve che non risulti ambigua: ad es. *A, però anche B potrebbe essere...*, e agevolare la valutazione automatica. Infatti il formato di questi risultati ne facilita la raccolta attraverso un semplice script che analizza la risposta data e la paragona alla risposta corretta di DiPRomIt-scenarios.

3.1.4 Le condizioni geografiche

Ciascuna istanza di DiPRomIt-scenarios è disponibile in due versioni: una contenente l'indicazione della regione di provenienza del proverbio corretto nel testo dello scenario (condizione *with-geo*, ✓) e una priva di tale informazione (condizione *without-geo*, ✗). Questo permette di misurare se e in quale misura i modelli sfruttino il segnale geografico quando devono abbinare uno scenario al proverbio appropriato.

3.1.5 Le metriche di valutazione

Per la valutazione si è scelto di utilizzare l'**accuracy** (proporzione di scenari per i quali il modello seleziona il proverbio corretto) e si riporta la distribuzione degli errori tra le categorie W_{same} , il modello sbaglia confondendosi con il proverbio della stessa regione di quello corretto, W_{close} , il modello sbaglia e sceglie un proverbio di una regione adiacente, W_{distant} , il modello sbaglia scegliendo un'opzione rappresentata da un proverbio di una regione

lontana da quella di partenza, e W_{other} , quando il modello restituisce un output non valido, ovvero non una delle lettere A, B, C, D. La distribuzione degli errori permette di rispondere a una domanda più articolata rispetto alla sola accuracy: non solo *quanto* sbaglia il modello, ma anche *come* sbaglia il modello. Per garantire la riproducibilità, ciascuna combinazione di modello, prompt e condizione geografica è eseguita con cinque seed diversi⁸ (0, 42, 101, 399, 565), per un totale di 15 esperimenti per modello per condizione e $147 \times 15 = 2.205$ istanze elaborate.

3.2 Presentazione dei risultati

Questa sezione presenta i risultati aggregati degli esperimenti descritti nella Sezione 3.1: la sottosezione seguente riporta le accuracy complessive per ciascun modello e condizione geografica, mentre la successiva ne discute i pattern principali; un'interpretazione più approfondita di questi risultati è rimandata al Capitolo 4.

3.2.1 Risultati complessivi

I risultati complessivi delle analisi su DiPRomIt-scenarios per tutti e dieci i modelli, nelle due condizioni geografiche, sono riportati in Tabella 3.2. I valori mostrati sono il risultato delle medie su cinque seed scelti in maniera casuale, per ciascuna delle tre varianti di prompt; i valori di accuracy della

⁸ Il *seed* determina il punto di inizio della randomizzazione; a parità di altri parametri, seed diversi possono produrre output diversi. L'uso di più seed permette di stimare la variabilità stocastica dei risultati.

condizione *with-geo* sono sottolineati per facilitare il confronto con la condizione *without-geo*.

Tabella 3.2: Risultati di DiPRomIt-scenarios per tutti i modelli con (✓) e senza (✗) informazione geografica. Accuracy media corredata di deviazione standard su 15 esperimenti per modello per condizione ($147 \times 15 = 2.205$ istanze totali processate). In grassetto i valori migliori assoluti per fascia parametrica. I valori sottolineati indicano l'accuracy nella condizione *with-geo*, per facilitarne il confronto con *without-geo*. Tra parentesi si possono notare i valori assoluti per tipologia di errore.

Modello	Geo?	Accuracy	W _{same}	W _{close}	W _{distant}	W _{other}
Llama-3.1-8B	✗	0.3519 $\pm$ 0.01	18.23% (402)	21.54% (475)	21.18% (467)	4.31% (95)
	✓	<u>0.3687$\pm$0.01</u>	18.10% (399)	21.27% (469)	20.14% (444)	3.63% (80)
Qwen-2.5-7B	✗	0.4572 $\pm$ 0.01	16.73% (369)	20.00% (441)	17.64% (389)	0.05% (1)
	✓	<u>0.4857$\pm$0.01</u>	16.46% (363)	19.77% (436)	15.19% (335)	0.00% (0)
Mistral-0.3-7B	✗	0.3206 $\pm$ 0.01	21.63% (477)	22.72% (501)	21.00% (463)	3.04% (67)
	✓	<u>0.3297$\pm$0.02</u>	20.14% (444)	22.90% (505)	20.73% (457)	3.27% (72)
Gemma-2-9B	✗	0.4485 $\pm$ 0.03	14.47% (319)	13.24% (292)	14.42% (318)	13.02% (287)
	✓	<u>0.4699$\pm$0.03</u>	14.15% (312)	12.83% (283)	13.56% (299)	12.47% (275)
Llamantino-3-ANITA-8B	✗	0.4272 $\pm$ 0.02	16.83% (371)	20.32% (448)	19.55% (431)	0.68% (15)
	✓	<u>0.4340$\pm$0.00</u>	19.37% (427)	19.50% (430)	17.51% (386)	0.68% (15)
Minerva-1.0-7B	✗	0.2730 $\pm$ 0.00	23.49% (518)	27.57% (608)	21.36% (471)	0.27% (6)
	✓	<u>0.2803$\pm$0.00</u>	23.22% (512)	27.39% (604)	21.18% (467)	0.18% (4)
Llama-3.1-70B	✗	0.5324 $\pm$ 0.02	17.46% (385)	16.92% (373)	11.70% (258)	0.68% (15)
	✓	<u>0.5437$\pm$0.02</u>	17.51% (386)	16.33% (360)	11.11% (245)	0.68% (15)
Qwen-2.5-72B	✗	0.6436 $\pm$ 0.00	13.06% (288)	11.29% (249)	11.29% (249)	0.00% (0)
	✓	<u>0.6617$\pm$0.00</u>	11.93% (263)	11.34% (250)	10.57% (233)	0.00% (0)
Mixtral-0.1-8x7B	✗	0.2113 $\pm$ 0.07	7.21% (159)	8.62% (190)	7.71% (170)	55.33% (1.220)
	✓	<u>0.2159$\pm$0.07</u>	6.76% (149)	8.98% (198)	7.03% (155)	55.65% (1.227)
Gemma-2-27B	✗	0.5206 $\pm$ 0.01	13.61% (300)	13.24% (292)	13.83% (305)	7.26% (160)
	✓	<u>0.5369$\pm$0.01</u>	12.56% (277)	13.24% (292)	13.38% (295)	7.12% (157)

3.2.2 Analisi dei pattern

I pattern complessivi che emergono dalla Tabella 3.2 sono ripresi nel dettaglio nel Capitolo 4; qui si commentano i principali elementi descrittivi.

L'informazione geografica fornisce un segnale utile per tutti i modelli (cfr. Tabella 3.2, valori sottolineati), sia di dimensioni ridotte che maggiori: ogni modello registra un'accuracy più alta nella condizione *with-geo* rispetto

a *without-geo*. Il guadagno medio è contenuto (+0,014 punti, pari a +1,4 punti percentuali), ma omogeneo su tutti i modelli senza eccezione, come confermato dal test di Wilcoxon signed-rank discusso nel Capitolo 4.

Tra i modelli di dimensioni ridotte con informazione geografica, Qwen-2.5-7B ottiene la performance migliore. Esso infatti seleziona il proverbio corretto in quasi la metà dei casi (0.4857), subito seguito da Gemma-2-9B (0.4699).

I modelli di dimensioni maggiori *with-geo* migliorano in generale i risultati della fascia ridotta appena discussa: il modello complessivamente migliore è Qwen-2.5-72B, che raggiunge 0.6617 di accuracy con deviazione standard pressoché nulla (0.00 se arrotondata a due decimali). Esso abbina correttamente quasi due terzi degli scenari.

La deviazione standard estremamente bassa di Qwen-2.5-72B rappresenta la solidità dei suoi risultati nel test: un valore così prossimo allo zero indica infatti che le prestazioni del modello variano pochissimo al variare del punto di inizio di randomizzazione generativa (seed), il che rende il valore di accuracy particolarmente solido. Qwen-2.5-72B inoltre, supera il secondo modello in classifica, Llama-3.1-70B (0.5437), di un margine non trascurabile (+0,118).

L'unico modello al di sotto della *random baseline* ($\approx 0,25$), cioè la soglia minima che permette di rifiutare l'opzione per la quale i modelli decidano completamente a caso, data dalla divisione di 1 su 4 scelte possibili, è Mixtral-0.1-8x7B. Mixtral-0.1-8x7B presenta problematicità in en-

trambe le condizioni, sia senza (0.2113) che con informazione geografica (0.2159). Una probabile causa è riscontrabile nell'elevatissima percentuale di output non validi: il modello non riesce a produrre una risposta nel formato atteso nel 55% circa dei casi (W_{other}).

Il problema potrebbe essere riconducibile al modo in cui il tokenizer di Mixtral segmenta i caratteri di risposta rispetto al vincolo di generazione: limitato ai nostri esperimenti a un singolo token, descritto in apertura di questo capitolo. Ad esempio, se per il tokenizer di Mixtral, «A» e « A» (con spazio iniziale), o un «a capo» al posto dello spazio, corrispondono a due token distinti, il vincolo potrebbe non includere la variante che il modello produrrebbe naturalmente.

Quasi tutti i modelli producono output non validi in misura trascurabile (0.00%–5.00%), con le eccezioni di Gemma-2-9B (12.47%–13.02%) e come già discusso Mixtral-0.1-8x7B. Per quanto riguarda la distribuzione degli errori tra W_{same} , W_{close} e W_{distant} , non emerge un andamento geograficamente motivato e sistematico: gli errori si distribuiscono in modo relativamente uniforme tra le tre categorie, suggerendo che i modelli non tendano sistematicamente a preferire distrattori geograficamente vicini al proverbio corretto.

3.3 Analisi degli errori

Questa sezione scende a un livello di analisi più granulare: scenari più facili e più difficili, effetto della condizione geografica, differenze tra i singoli

modelli.

3.3.1 Scenari più facili e più difficili

La Figura 3.1 riporta i 15 scenari più facili e i 15 più difficili per accuracy media aggregata su tutti i modelli, prompt, seed e condizioni geografiche. Abbiamo scelto 15 scenari perché rappresentano circa il 10% dei 147 scenari di DiPRomIt-scenarios⁹.

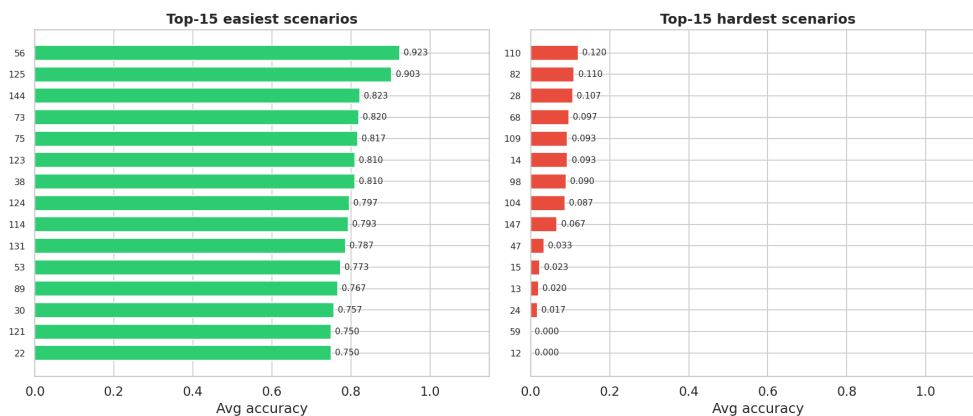


Figura 3.1: Top- k ($k = 15$) scenari globalmente più facili (*sinistra*) e più difficili (*destra*) per accuracy media aggregata su tutti i modelli, prompt, seed e condizioni geografiche.

L'ispezione manuale degli scenari più facili rivela che almeno tre di essi (ID 38, 53, 73) sono costruiti attorno a proverbi di uso diffuso in italiano. Ad esempio, lo scenario ID 38 è associato al proverbio friulano «*Quand che il giat no l'è in paīs, a fās fieste la surīs*», corrispondente all'italiano «*Quando il gatto non c'è i topi ballano*». Questo dato è compatibile con l'ipotesi che i proverbi che hanno *viaggiato attraverso le lingue*, ovvero che possiedono un equivalente consolidato, risultino più facilmente memorizzati dai LLM e

⁹ La scelta di $k = 15$ come soglia di analisi rispecchia quella adottata nell'articolo di riferimento (Errico et al., 2026, § 4.3), al fine di mantenere la comparabilità dei risultati.

quindi più rapidamente identificabili: un percorso che rifletterebbe, più che un'autentica sensibilità culturale, un'inferenza riconducibile alla memorizzazione. Si tratta tuttavia di un'osservazione qualitativa condotta su soli tre scenari. Per ampliare questo filone d'ipotesi potrebbe essere utile, seppur non risolutiva da sola, una stima della frequenza con cui ciascun proverbio (o il suo equivalente) compare nel pre-addestramento, oppure variabilizzare il problema con un confronto tra scenari con e senza equivalenti noti. Tale verifica è lasciata a lavori futuri (cfr. Capitolo 5).

Per gli scenari più difficili, al contrario, non si riscontra alcun equivalente italiano noto per la maggior parte dei proverbi associati. Il caso del proverbio «*Il vino è la mammella dei vecchi*» (ID 372) costituisce però un controesempio rilevante per l'ipotesi formulata nel paragrafo precedente: pur essendo attestato in sei varietà diverse come riportato in DiPRomIt, e quindi tutt'altro che isolato, risulta sistematicamente difficile per tutti i modelli.

Questo caso suggerisce che la sola diffusione di un proverbio (o di un suo equivalente) tra le varietà non sia condizione sufficiente per un'identificazione corretta, e che altri fattori, come la salienza, o frequenza d'uso, debbano intervenire per il corretto riconoscimento. Nel complesso la località di riferimento culturale emerge dunque come il fattore principale che determina la difficoltà di uno scenario.

3.3.2 Effetto della condizione geografica

La Figura 3.2 disaggrega la stessa analisi separando le due condizioni geografiche.

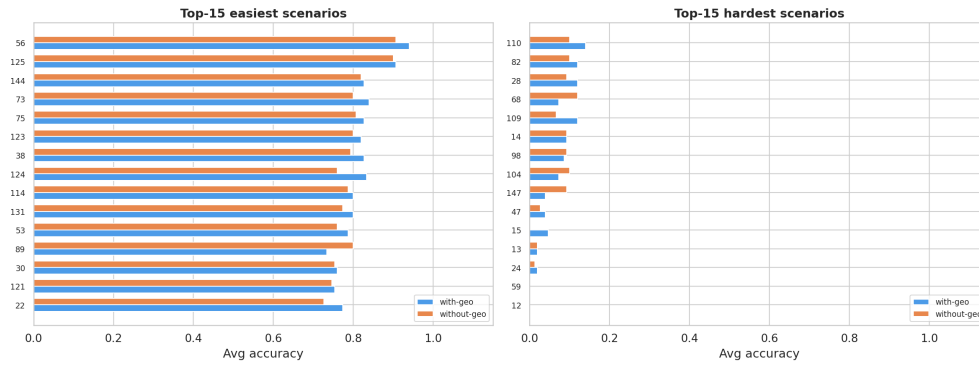


Figura 3.2: Top- k ($k = 15$) scenari più facili (*sinistra*) e più difficili (*destra*) disaggregati per condizione geografica (*with-geo* ✓ e *without-geo* ✗).

Tra gli scenari più facili, l'informazione geografica si rivela chiaramente utile: in 14 casi su 15, la condizione *with-geo* produce un'accuracy superiore rispetto a *without-geo*. Per gli scenari più difficili il quadro è invece meno netto: l'informazione geografica è utile, ma solo nella quasi metà dei casi (7 su 15); nei restanti 8 scenari, invece, l'informazione geografica produce un effetto negativo, come discusso nel Capitolo 4, presumibilmente perché il segnale regionale induce i modelli a privilegiare il distrattore W_{same} rispetto al proverbio corretto.

3.3.3 Analisi per modello

Infine, la Figura 3.3 estende l'analisi al livello del singolo modello, riportando l'accuracy media per modello sui top- k scenari più facili e più difficili.

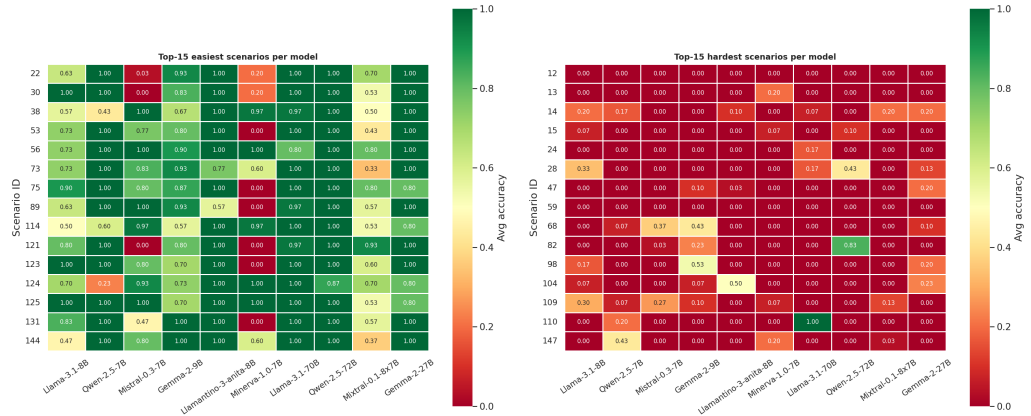


Figura 3.3: Heatmap che mostra l'accuracy media per modello sui top- k ($k = 15$) scenari più facili (*sinistra*) e più difficili (*destra*), aggregata di tutte le condizioni sperimentali.

Come si nota dalla Figura 3.3 i modelli di dimensioni maggiori mostrano un'accuracy generalmente più alta in entrambe le sotto-popolazioni di scenari, con quelli di dimensioni ridotte sistematicamente al di sotto. Il divario è particolarmente pronunciato con gli scenari difficili, dove anche i modelli più performanti raggiungono accuracy molto basse.

Per riassumere le categorie di errori, basti pensare che le categorie W_{same} , W_{close} e W_{distant} costituiscono *errori di contenuto*: il modello produce un output valido, ma seleziona il proverbio sbagliato. W_{other} costituisce invece un *errore di formato*: non viene prodotta nessuna risposta valida.

Il quadro complessivo sarà discusso e interpretato nel Capitolo 4.

Capitolo 4

Discussione

Il capitolo interpreta i risultati sperimentali presentati nel Capitolo 3 alla luce delle domande di ricerca formulate nell'Introduzione, che hanno guidato il lavoro. Salvo diversa indicazione, i valori di accuracy discussi nelle sezioni seguenti si riferiscono alla condizione *with-geo*; il vantaggio sistematico di questa condizione rispetto a *without-geo* è illustrato nella Sezione 4.4. Il capitolo si articola in sei sezioni. La Sezione 4.1 presenta il task a scelta multipla, i limiti nell'interpretazione dei risultati e i rischi legati alla crescente tendenza di una lettura antropomorfizzante del comportamento dei modelli. La Sezione 4.2 discute il rapporto tra scala parametrica e performance. La Sezione 4.3 analizza il comportamento dei modelli specializzati sull'italiano. La Sezione 4.4 approfondisce il ruolo dell'informazione geografica. La Sezione 4.5 argomenta l'ipotesi dei proverbi itineranti, qui chiamati «proverbi che hanno viaggiato», come fattore predittivo dei risultati. La Sezione 4.6 riprende il caso di studio del gruppo VINO → LATTE come illustrazione concreta delle dinamiche osservate a livello aggregato.

4.1 Natura del task

Prima di analizzare nel dettaglio i risultati, è necessario chiarire la natura del task e i limiti che ne derivano per la sua interpretazione, dalla soglia minima di accuratezza attesa per puro caso al rischio di letture antropomorfizzanti.

4.1.1 Baseline casuale

Le prestazioni calcolate oscillano tra il 25% (random baseline) e il 66% (punteggio massimo ottenuto) di accuratezza. Uno dei dieci modelli presi in considerazione, Mixtral-0.1-8x7B, è stato escluso dai confronti comparativi tra modelli e famiglie (Sezione 4.2) poiché rimane eccessivamente vicino alla random baseline, o al di sotto di essa; resta invece incluso nelle analisi successive, come l'effetto dell'informazione geografica (Sezione 4.4).

La motivazione dell'esclusione è legata alla natura stessa del task: in un test a scelta multipla con quattro opzioni, un modello privo di qualsiasi conoscenza dell'argomento raggiungerebbe statisticamente un'accuratezza di circa il 25% rispondendo in modo completamente casuale (Myrzakhan et al., 2024). Questo fenomeno, definito come «random guessing» o «lottery ticket choice», indica che il modello può indovinare la risposta corretta per puro caso.

Le ricerche correnti evidenziano infatti come molti modelli recenti presentino prestazioni prossime alla casualità (*near random-chance accuracy*) su svariati benchmark, rendendo necessario che i risultati superino signifi-

cativamente tale soglia per dimostrare una reale capacità di comprensione e problem solving (Hendrycks et al., 2021).

4.1.2 Cautele interpretative

Per gli altri modelli l'interpretazione rimane valida, sebbene non sia possibile escludere del tutto l'ipotesi che alcune risposte siano effettivamente falsi positivi. Questa ipotesi appare poco plausibile alla luce del fatto che nove modelli su dieci hanno ottenuto punteggi validi. Tale considerazione è supportata inoltre dall'utilizzo di cinque seed identici per la randomizzazione delle risposte.

L'impiego di molteplici seed di generazione rappresenta una strategia consolidata per ridurre l'influenza della componente stocastica introdotta durante il processo di generazione. Un aspetto spesso trascurato dai benchmark che si affidano a campionamenti singoli o deterministici (Zhang et al., 2025). Sebbene i testi analizzati nella letteratura non indichino esplicitamente l'uso di cinque ripetizioni come standard, si ritiene che la quantificazione data sia sufficiente per un calcolo statistico significativo della varianza (Madaan et al., 2024).

In breve, i valori di accuracy andrebbero trattati più come «indici» che come misure esatte: rimane una limitata possibilità, ritenuta poco probabile, ma non escludibile, che parte delle prestazioni osservate sia stata influenzata dalla componente stocastica del processo di generazione, con un conseguente lieve sovradimensionamento dei valori rispetto alle reali capacità del

modello. Una limitazione da tenere presente nell'interpretazione dei risultati. Anche nell'interpretazione più favorevole, il miglior modello raggiunge comunque solo il 66% di accuracy, un risultato discreto ma non elevato; tutti gli altri modelli si fermano più in basso. Alla luce di queste limitazioni, appare opportuna una lettura prudente e critica dei risultati, piuttosto che ottimistica.

4.1.3 Il rischio dell'antropomorfizzazione

Un'analogia utile per ricontestualizzare i risultati può essere tratta dai test a scelta multipla somministrati agli studenti, che seguono la stessa logica. Qualora uno studente rispondesse completamente a caso, questo potrebbe teoricamente ottenere un voto elevato, ad esempio 27 o 28 su 30 (voto universitario), ma la probabilità che ciò accada risulta realisticamente molto ridotta (Frary, 1988). Naturalmente, questa analogia presenta importanti limitazioni. Nel caso di un partecipante umano, infatti, è raro riscontrare uno stato di impreparazione totale; gli esaminandi utilizzano spesso conoscenze parziali per eliminare i distrattori o si affidano a intuizioni e «sensazioni» che portano a successi superiori a quelli del puro caso (*ibidem*). Usare gli stessi termini psicologici impiegati per l'analogia con gli studenti, come «intuizioni» e «conoscenza parziale», per descrivere il comportamento del modello si può considerare un campanello d'allarme: è necessario evitare di antropomorfizzare i LLM. Ciò che l'analogia cerca di dimostrare è un fatto quantitativo, valido a prescindere da chi produce le risposte: uno scarto

sistematico dalla pura casualità implica che un qualche segnale, oltre al caso, guidi le scelte. Nel caso umano quel segnale assume una forma diversa: in un LLM esso coincide con la previsione statistica del token successivo più probabile.

Sebbene la fluidità del linguaggio generi una potente «illusione dell'utente linguistico» che spinge a credere, e quindi descrivere, il loro comportamento tramite categorie umane, tali modelli non possiedono intenzioni comunicative (Shanahan, 2024). Di conseguenza, è un «serio errore» trasferire direttamente ai modelli i presupposti che caratterizzano la valutazione di un soggetto umano, poiché il loro funzionamento è basato sulla previsione statistica di token (*ibidem*). Ad esempio, si è evitato di introdurre istruzioni quali «se non conosci la risposta, dichiara di non saperla» o «pensaci bene» o prevedere penalizzazioni: i LLM mancano di un accesso alla verità esterna e di stati mentali quali l'onestà o la reticenza. A causa della loro «fragilità» e sensibilità alle variazioni del prompt, tali indicazioni modificherebbero l'output in modo difficilmente prevedibile. Per questo motivo, si è scelto di mantenere un protocollo di valutazione il più possibile uniforme e controllato.

4.2 Scala parametrica e performance

Riprendendo la Tabella 3.2, il primo risultato che emerge è la scarsa influenza che la scala parametrica ha nel costituire un predittore lineare della performance su DiPRomIt-scenarios.

4.2.1 Confronto tra famiglie di modelli

A parità di famiglia di modelli, il guadagno del modello maggiore rispetto al minore è asimmetrico: Qwen-2.5-72B supera Qwen-2.5-7B di circa 18 pp (0.6617 vs. 0.4857 con informazione geografica), mentre Gemma-2-27B supera Gemma-2-9B di 7 pp (0.5369 vs. 0.4699), e Llama-3.1-70B supera Llama-3.1-8B di quasi 18 pp (0.5437 vs. 0.3687). La relazione scala/performance non è dunque costante tra famiglie, suggerendo che i dati di addestramento e le scelte architetturali abbiano un peso almeno comparabile al numero di parametri.

Tabella 4.1: Guadagno di accuracy tra modello minore e maggiore, a parità di famiglia (variante *with-geo*). La riga evidenziata in giallo rappresenta la famiglia con il guadagno minore tra versione piccola e versione grande.

Famiglia	Modello (piccolo)	Acc.	Modello (grande)	Acc.	
Qwen	Qwen-2.5-7B	0.4857	Qwen-2.5-72B	0.6617	+17.6
Llama	Llama-3.1-8B	0.3687	Llama-3.1-70B	0.5437	+17.5
Gemma	Gemma-2-9B	0.4699	Gemma-2-27B	0.5369	+6.7

In sintesi, scalare da 9B a 27B dentro la famiglia Gemma produce un miglioramento molto più modesto di quanto avvenga scalando dentro le famiglie Qwen o Llama.

Un secondo elemento di rilievo riguarda le inversioni inter-famiglia: casi in cui, a parità di dimensione, un modello di una famiglia supera nettamente quelli di altre famiglie, contraddicendo l'aspettativa che la performance sia determinata principalmente dalla scala. Gemma-2-9B (9B parametri) ne è un esempio, con 0.4699 di accuracy, che supera sia Llama-3.1-8B (0.3687)

che Mistral-0.3-7B (0.3297). A parità di scala il divario inter-famiglia risulta marcato, come si evince dalla Tabella 4.2.

Tabella 4.2: Confronto inter-famiglia 7–9B. In giallo il modello in prima posizione.

Pos.	Modello	Parametri	Accuracy (<i>with-geo</i>)
1°	Gemma-2-9B	9B	0.4699
2°	Llama-3.1-8B	8B	0.3687
3°	Mistral-0.3-7B	7B	0.3297

4.2.2 Quadro riassuntivo di tutti i modelli

Per completare il quadro con una visione d'insieme su tutti i modelli valutati, le Tabelle 4.3 e 4.4 riportano la classifica completa dei modelli, rispettivamente, di dimensioni ridotte e maggiori, nella condizione *with-geo*¹.

Tabella 4.3: Classifica dei modelli di dimensioni ridotte (7–9B), condizione *with-geo*. In giallo il modello in prima posizione.

Pos.	Modello	Parametri	Accuracy
1°	Qwen-2.5-7B	7B	0.4857
2°	Gemma-2-9B	9B	0.4699
3°	Llamantino-3-ANITA-8B	8B	0.4340
4°	Llama-3.1-8B	8B	0.3687
5°	Mistral-0.3-7B	7B	0.3297
6°	Minerva-1.0-7B	7B	0.2803

Tabella 4.4: Classifica dei modelli di dimensioni maggiori (27–72B), condizione *with-geo*. In giallo il modello in prima posizione.

Pos.	Modello	Parametri	Accuracy
1°	Qwen-2.5-72B	72B	0.6617
2°	Llama-3.1-70B	70B	0.5437
3°	Gemma-2-27B	27B	0.5369

¹ Per i valori completi, incluse le condizioni *without-geo* e la distribuzione degli errori, cfr. Tabella 3.2 nel Capitolo 3.

Fa eccezione al quadro generale Mixtral-0.1-8x7B, il cui risultato è al di sotto della baseline casuale ($\approx 0,25$). Il risultato è probabilmente attribuibile a un problema di formato di output: nel 55% circa dei casi il modello produce un output non valido (W_{other}). La causa soggiacente potrebbe essere le istruzioni troppo «restrittive», ma questo non spiegherebbe il perché delle altre risposte plausibili. Appurato che il modello abbia risposto in una misura accettabile, è stato deciso che il comportamento più corretto fosse quello di escluderlo dal confronto diretto e considerarlo un caso limite. Complessivamente, il quadro deducibile è insufficiente: fissando la soglia di sufficienza al 60% di accuracy, solo Qwen-2.5-72B la raggiunge (66,2%), tutti gli altri nove modelli si fermano tra il 21,6% (Mixtral-0.1-8x7B) e il 54,4% (Llama-3.1-70B).

4.3 I modelli specializzati sull'italiano

I modelli pre-addestrati su dati testuali in italiano presentano dei risultati alquanto controintuitivi. L'ipotesi di partenza per la quale una maggiore esposizione all'italiano, lingua-tetto dei dialetti romanzi d'Italia, potesse favorire la comprensione culturale dialettale, è smentita dai risultati presentati qui.

Minerva-1.0-7B è il modello con la performance più bassa tra tutti quelli valutati (0.2730 senza informazione geografica, 0.2803 con informazione geografica), appena al di sopra della random baseline. Llamantino-3-ANITA-8B ottiene risultati discreti, ma insufficienti (0.4272/0.4340),

senza particolari vantaggi apprezzabili rispetto a modelli generici, come Qwen-2.5-7B.

Un'ultima considerazione riguarda l'analisi delle distribuzioni delle risposte fornite dai modelli. In linea teorica sarebbe possibile analizzare la frequenza con cui ciascun modello seleziona le diverse opzioni (A, B, C o D) per identificare eventuali pattern ricorrenti; questo tipo di analisi richiede però particolare cautela interpretativa.

Una preferenza sistematica per una determinata opzione, come quella piuttosto marcata mostrata da Minerva-1.0-7B verso l'opzione A (84,4% delle risposte in condizione with-geo), non può essere automaticamente interpretata come evidenza di un bias specifico: si tratterebbe in ogni caso di un'osservazione descrittiva, dalla quale non è possibile risalire con certezza, in modo causale e verificabile, ai meccanismi interni che la generano (Tang et al., 2026; Zheng et al., 2024) — che si tratti di regolarità presenti nei dati di addestramento o di altri fattori non osservabili del processo di training. Tuttavia, data l'enorme scala di tali dati, del resto, una verifica diretta di queste ipotesi risulterebbe impraticabile.

Il caso di Minerva-1.0-7B merita una nota a parte. L'accuratezza molto bassa, unita a un'apparente uniformità dei risultati tra le esecuzioni, ha portato alla necessità di un'ispezione manuale delle risposte grezze del modello: un controllo mirato a questa specifica anomalia, non esteso sistematicamente agli altri nove modelli valutati. L'ispezione rivela che Minerva-1.0-7B seleziona la lettera **A** nell'84,4% delle 2.205 istanze

processate in condizione *with-geo* (1.860 casi su 2.205), a fronte di una distribuzione della posizione della risposta corretta relativamente equilibrata tra i quattro slot (A: 28,6%, B: 20,4%, C: 19,7%, D: 31,3%, fissa per scenario grazie allo shuffling descritto nel Capitolo 2). Su un'accuracy complessiva di appena 0.2803 (28,03% delle 2.205 istanze), 530 delle 618 risposte corrette del modello corrispondenti, cioè all'85,8% dei (pochi) casi in cui il modello ha scelto correttamente sarebbero spiegabili semplicemente dal fatto che la lettera A corrispondeva per coincidenza al proverbio corretto: l'accuracy di Minerva-1.0-7B è dunque in larghissima parte un artefatto della posizione di scelta delle risposte e della preferenza del modello.

Questa verifica solleva una domanda che il presente lavoro lascia aperta: se un pattern analogo fosse presente, anche in forma più lieve, in uno o più degli altri modelli valutati, l'accuracy riportata in Tabella 3.2, e di conseguenza le classifiche presentate nella Sezione 4.2.2, potrebbero non riflettere fedelmente la competenza culturale dei modelli, ma in parte essere influenzate da un artefatto posizionale non ancora quantificato. In ottica di completezza, un'estensione sistematica di questa analisi a tutti i modelli costituirebbe una verifica importante.

4.4 Il segnale geografico: utile ma non sufficiente

Tutti e dieci i modelli ottengono un'accuracy più alta nella condizione *with-geo* rispetto a *without-geo* (cfr. Tabella 3.2 nel Capitolo 3), come si può notare osservando il guadagno medio di accuracy, che, passando

da *without-geo* a *with-geo*, è di +1,4 punti percentuali. La consistenza universale dell'effetto, con nessuna eccezione tra i dieci modelli, suggerisce che l'informazione geografica costituisca un segnale utile per il corretto riconoscimento della risposta.

Per verificarne la significatività statistica, è stato condotto un test di Wilcoxon signed-rank² sulle coppie di accuracy per modello nelle due condizioni. Il test restituisce $p = 0,002$, confermando che la differenza è statisticamente significativa al livello $\alpha = 0,05$ (Dror et al., 2020). Per la verifica delle assunzioni del test vedere la tabella dei dati appaiati e il calcolo passo-passo, cfr. Appendice C.

Tuttavia, la modestia del guadagno medio (+1,4 pp) pone la questione di come i modelli stiano effettivamente usando l'informazione geografica. Sono plausibili tre ipotesi interpretative: (i) un effetto di *disambiguazione lessicale*, per cui, quando il testo dello scenario specifica la regione, aumenta la probabilità assegnata dal modello al token corrispondente, avendo fatto precedentemente scattare un'associazione di correttezza dalle caratteristiche lessicali/dialettali del testo del proverbio in ciascuna opzione, ha disambiguato lessicalmente; (ii) un effetto di *ancoraggio culturale*, per cui la menzione della regione potrebbe attivare rappresentazioni culturali associate a quell'area (cucina, usi, paesaggio, personaggi tipici), che facilitano indirettamente il riconoscimento del proverbio pertinente; (iii) un effetto distorsivo, che emerge chiaramente nell'analisi degli scenari difficili

² Il test di Wilcoxon signed-rank è una versione non parametrica del t-test su campioni appaiati, appropriata quando la normalità della distribuzione non può essere assunta, come nel caso di campioni di dimensioni ridotte ($n = 10$).

(Figura 3.2), per cui l'informazione geografica può indurre il modello a selezionare il distrattore W_{same} , proveniente dalla stessa regione del proverbio corretto, invece del proverbio semanticamente pertinente.

In 8 casi su 15 tra gli scenari più difficili, la condizione *with-geo* produce risultati peggiori rispetto a *without-geo*. In altri termini, il modello apprende a usare il segnale regionale come proxy per la risposta corretta, ma questo segnale è fuorviante: la regione è condivisa da molti proverbi (inclusi i distrattori W_{same}), e solo la comprensione semantica del contesto permette di distinguere il proverbio corretto dagli altri.

In sintesi (cfr. Sezione 2.3), per i modelli «vedere» l'informazione geografica aiuta, ma il guadagno non è tuttavia legato alla scala parametrica, sostanzialmente equivalente tra modelli piccoli e grandi (rispettivamente 1,50 e 1,52 pp in media, cfr. Tabella 3 in Appendice C); a beneficiarne meno sono piuttosto i due modelli specializzati sull'italiano, Llamantino-3-ANITA-8B e Minerva-1.0-7B (0,70 pp in media, contro 1,74 pp per gli altri modelli).

4.5 Proverbi che hanno viaggiato: memorizzazione vs. sensibilità culturale

L'analisi degli scenari estremi (Figura 3.1) permette di formulare un'ipotesi interpretativa con rilevanza teorica più ampia: la difficoltà di uno scenario tende a diminuire all'aumentare del grado di radicamento locale del proverbio sottostante.

Gli scenari più facili sono costruiti attorno a proverbi che possiedono un equivalente consolidato in italiano standard o che appartengono alla paremiologia pan-europea. Il proverbio friulano, associato allo scenario ID 38, «*Quand che il giat no l'è in paīs, a fās fieste la surīs*», ha un corrispondente in italiano: «*Quando il gatto non c'è i topi ballano*», che ha anche una versione corrispondente inglese «*When the cat's away, the mice will play*». In questo caso possiamo parlare di proverbio «ben viaggiato», o «itinerante», cioè un enunciato particolarmente conosciuto a livello pan-europeo: la sua diffusione è attestata in 46 lingue europee (Mieder, 2004, p. 12). Si sospetta che proverbi di questo genere siano particolarmente semplici perché più rappresentati nei dati di training (Kandpal et al., 2023): una più alta accuracy potrebbe quindi semplicemente riflettere un effetto di memorizzazione, piuttosto che un'autentica «comprensione» dello scenario.

Gli scenari più difficili, al contrario, sono associati a proverbi che non hanno un equivalente italiano noto e che esprimono frame concettuali localizzati, spesso legati a pratiche, oggetti o situazioni caratteristici di una specifica area geografica o varietà dialettale. Per questi proverbi, il modello non può affidarsi a conoscenze generali acquisite durante il pre-addestramento: deve inferire la pertinenza culturale a partire dal testo dello scenario e dalle opzioni proposte, operazione che richiede una forma di ragionamento culturale che gli LLM attuali non dimostrano di padroneggiare.

Questa distinzione richiama il costrutto teorico della *graded salience* (Giora, 1997): espressioni idiomatiche o proverbiali con alta salienza (per-

ché frequenti, diffuse o standardizzate) sono processate più rapidamente e con maggiore affidabilità rispetto a quelle con bassa salienza.

4.6 Caso di studio: il gruppo VINO → LATTE

Il gruppo di varianti 372 (cfr. Tabella 2.6), composto da sei proverbi che condividono il frame concettuale VINO → LATTE in altrettante varietà dialettali, costituisce un caso di studio privilegiato perché consente di isolare il contributo della trasparenza semantica dal fattore confondente della distanza geografica.

Come introdotto nella Sezione 2.3, la variante friulana (ID 434) «*El vin al è latt da puars vecios*» risulta essere la più facile tra le sei in quasi tutti i modelli.

Una possibile ipotesi è il trasferimento cross-linguistico dei «proverbi viaggiati». Siccome gli LLM apprendono da uno spazio di embedding condiviso tra più lingue, i concetti semanticamente simili finiscono più vicini tra loro che i loro corrispettivi con meno connessioni. Questo fenomeno avverrebbe indipendentemente dalla lingua di input, fenomeno chiamato anche «lingua latente» (M. Wang, Adel et al., 2025), attraverso la maggior parte dei loro strati (*layer*). In questo spazio condiviso i modelli effettuerebbero un transfer interlinguistico (Lu et al., 2018), prendendo una scorciatoia, diversa da quella vista precedentemente, che gli permetterebbe di raggiungere una risposta plausibile più facilmente rispetto alle domande con scenari non rappresentati sufficientemente nello spazio condiviso. Se

accettassimo questa ipotesi non si spiegherebbe il fenomeno di Qwen-2.5-72B, che ottiene buoni risultati, ma mostra una performance bassa con la variante friulana in questione.

La variante lombarda (ID 372) è invece sistematicamente la più difficile tra le sei, con accuracy prossima allo zero in entrambe le condizioni geografiche per quasi tutti i modelli. Questo scenario figura tra i dieci più difficili dell'intero DiPRomIt-scenarios.

In questo gruppo, la combinazione di opacità lessicale equivale all'assenza di rappresentazioni vettoriali equivalenti. Il frame concettuale implicito crea una barriera interpretativa che anche il modello migliore non supera facilmente. Nel complesso, il gruppo VINO → LATTE esemplifica le dinamiche osservate a livello aggregato (cfr. Sezione 4.5).

Capitolo 5

Conclusioni

Il presente capitolo chiude il lavoro, riprendendo quanto emerso nei capitoli precedenti alla luce delle domande di ricerca formulate nell’Introduzione.

Il capitolo si articola in tre sezioni. La Sezione 5.1 riassume i principali risultati per ciascuna. La Sezione 5.2 esamina criticamente i limiti della risorsa e degli esperimenti condotti. La Sezione 5.3 propone infine alcune direzioni di ricerca future.

5.1 Sintesi del contributo

Questo lavoro ha presentato due nuove risorse per il NLP culturale: DiPRomIt, il primo dataset di proverbi scritti in varietà romanze dialettali d’Italia, e DiPRomIt-scenarios, un benchmark per la valutazione della sensibilità culturale *contestuale* dei LLM, su cui sono stati valutati 10 modelli open-weight. Dai risultati sperimentali emergono i seguenti riscontri principali, discussi in relazione alle domande di ricerca poste nell’Introduzione:

RQ1 «I modelli linguistici abbinano correttamente i proverbi allo scenario appropriato, oppure la possibile incompetenza è riconducibile a una

causa, in questo caso a meccanismi di memorizzazione?»

I modelli abbinano correttamente i proverbi allo scenario appropriato solo in parte (il modello migliore raggiunge circa due terzi di accuratezza). La difficoltà di uno scenario è correlata, in primo luogo alla sua diffusione 4.5 e in secondo luogo all'alto valore paremiaco 1.1.

La ragione va cercata nella memorizzazione.

RQ2 «L'inclusione esplicita di informazioni sulla provenienza geografica nel prompt fornisce ai modelli un segnale utile per l'abbinamento corretto del proverbio?»

Sì, ma in modo disomogeneo: aiuta in modo netto negli scenari più facili dove guida la risposta corretta in quattordici casi su quindici; negli scenari più ostici, invece, il suo contributo si dimezza, risultando utile in appena sette casi su quindici.

RQ3 «Quali pattern emergono dall'analisi degli errori dei modelli, in particolare rispetto a quanto appena descritto?»

Il pattern più chiaro riguarda il comportamento dei modelli quando il segnale geografico non basta a orientare la risposta: in questi casi il modello tende a preferire il distrattore più vicino W_{same} . A livello di confronto tra modelli, la scala parametrica non risulta un predittore lineare della performance (la qualità dei dati di pre-addestramento e le scelte architetturali incidono) e i modelli specializzati sull'italiano (Minerva, Llamantino) non mostrano particolari vantaggi rispetto ai modelli generici.

RQ4 «I modelli mostrano una forma di conoscenza dei proverbi emergente, paragonabile a quella della gente comune?»

No, i risultati inducono a rifiutare un’ipotesi secondo la quale i modelli mostrano una forma di conoscenza dei proverbi emergente, paragonabile a quella della gente comune. In teoria questa domanda trova parziale risposta nell’asimmetria osservata nella prima domanda. L’accuratezza si alza con proverbi con equivalente diffuso, e dunque, più rappresentati nei dati di pre-addestramento.

DiPRomIt e DiPRomIt-scenarios sono rilasciati pubblicamente per favorire la ricerca all’intersezione tra linguistica, paremiologia e NLP culturale¹ (Errico et al., 2026).

5.2 Limitazioni

Accanto ai contributi appena descritti, è opportuno considerare con altrettanta chiarezza i limiti che ne circoscrivono la portata, distinti tra quelli propri della risorsa e quelli specifici del disegno sperimentale.

5.2.1 Limitazioni della risorsa

DiPRomIt è costruito a partire da WikiQuote, una raccolta di citazioni a contribuzione volontaria. Benché il sito fornisca linee guida editoriali, la loro applicazione uniforme non può essere garantita: alcune trascrizioni, traduzioni o attribuzioni potrebbero essere imprecise o difficilmente verifi-

¹ <https://github.com/dhfbk/dialectal-proverbs>

cabili di prima mano. La verifica a campione, che è stata condotta durante la raccolta (cfr. Sezione 2.1.1), ha mitigato questo rischio.

Una porzione significativa delle fonti bibliografiche associate ai proverbi è stata esclusa per indisponibilità in catalogo o mancata digitalizzazione delle fonti. Questo ha reso impossibile la verifica sistematica della corrispondenza tra il testo su WikiQuote e la fonte originale; alcune voci sono state controllate a campione tramite Google Libri o link disponibili all'epoca della consultazione.

Va infine segnalato che la categorizzazione dei proverbi per regione amministrativa costituisce una semplificazione operativa. I dialetti romanzi d'Italia si dispongono su un continuum e le isoglosse che li caratterizzano attraversano spesso i confini amministrativi regionali (cfr. Berruto & Cerruti, 2015): il siciliano è parlato anche nella Calabria meridionale, il friulano coesiste con il veneto in Friuli-Venezia Giulia, il ladino è presente in Trentino-Alto Adige (Ramponi, 2024). L'adozione della regione come unità di analisi è giustificata da ragioni pratiche, ma non corrisponde alla reale struttura della variazione dialettale.

5.2.2 Limitazioni degli esperimenti

Gli esperimenti si limitano a un regime zero-shot. Le impostazioni few-shot potrebbero migliorare le performance, ma non sono state contemplate in questa ricerca. Questa scelta è stata dettata da ragioni computazionali e di comparabilità, ma non permette di valutare la capacità di adattamento

rapido dei modelli a contesti culturali nuovi.

Il set di modelli valutato, per quanto rappresentativo dell’offerta open-weight alla data degli esperimenti, potrebbe beneficiare di maggiore copertura. In particolare, non sono stati valutati modelli di ragionamento (reasoning LLM), come quelli basati su tecniche di chain-of-thought o reinforcement learning from human feedback² orientato alla riflessione multi-step, che potrebbero trarre vantaggio dalla struttura argomentativa implicita nei proverbi.

Come metrica di valutazione è stata adottata l’accuracy in un task a scelta multipla. Tuttavia questa scelta non cattura appieno il percorso induttivo che porta il modello a scegliere una risposta, né distingue tra risposte corrette raggiunte tramite meccanismi di scorciatoia (ad esempio indizi lessicali superficiali o pattern statistici nella formulazione delle opzioni). Un’analisi più raffinata del comportamento del modello, come l’analisi della distribuzione di probabilità sulle opzioni o la raccolta di spiegazioni generate dal modello, potrebbe offrire una finestra più ricca sulla natura del processo inferenziale. Un caso concreto di questo limite emerge dall’analisi di Minerva-1.0-7B (cfr. Sezione 4.3): l’84,4% delle sue risposte in condizione *with-geo* coincide con la lettera A, mentre la posizione della risposta corretta è distribuita in modo pressoché equilibrato tra le quattro opzioni; l’85,8% delle sue risposte corrette è quindi una coincidenza.

² Reinforcement Learning from Human Feedback (apprendimento per rinforzo da feedback umano): tecnica di addestramento supplementare in cui il modello viene ottimizzato sulla base di giudizi umani sulla qualità delle proprie risposte, per allinearne il comportamento alle preferenze umane.

Il presente lavoro non riporta stime del costo computazionale, energetico o idrico degli esperimenti, un dato che la comunità dovrebbe considerare rilevante e che contribuirebbe a creare standard per lavori futuri.

5.3 Prospettive future

Il presente lavoro apre diverse direzioni di ricerca, tanto sul piano della risorsa quanto su quello del benchmark e degli esperimenti.

Ampliamento delle risorse Una prima direzione riguarda il completamento delle porzioni di dataset ridotte per vincoli di tempo o lunghezza.

Alcune sezioni, in particolare quelle calabresi, napoletane e tarantine, sono state limitate a 50 proverbi ciascuna, e alcune fonti WikiQuote sono state esplorate solo parzialmente. Risulta quindi naturale una loro estensione consentirebbe anche una copertura più equilibrata. Parallelamente, alcune regioni sono state escluse per la scarsità di fonti verificabili: la ricerca di risorse attestate colmerebbe la sottorappresentazione delle regioni con meno di 10 proverbi (ed es. Marche, Toscana, Trentino-Alto Adige) e renderebbe possibile l'inclusione di regioni come Umbria e Valle d'Aosta.

Sul piano della copertura linguistica, DiPRomIt include esclusivamente varietà italo-romanze. Un ampliamento naturale riguarderebbe le altre varietà del territorio italiano, come il francoprovenzale (patois valdostano), il ladino dolomitico e il tedesco sudtirolese e altre varietà gallo-romanze o retoromaniche presenti in aree di confine.

Un'estensione più ambiziosa riguarda infine la dimensione pan-romanza: la raccolta di proverbi equivalenti in altre tradizioni romanze europee (occitano, catalano, galego, romeno) permetterebbe di costruire un atlante paremiologico computazionale a scala continentale, rendendo esplicita la rete di variazione diatopica oltre i confini nazionali.

Sul piano del benchmark, due direzioni sono prioritarie. La prima è di natura computazionale: attualmente gli agenti narrativi presenti negli scenari (personaggi, ruoli, relazioni familiari) sono scritti tra marcatori strutturati — ad esempio tra parentesi quadre, come in [nonna], [bambina], [venditore] — in origine pensati per consentire un rimescolamento (*shuffling*) automatico dei ruoli. Come discusso in Sezione 2.2.2, dopo qualche tentativo l'idea è stata accantonata: non si riusciva a gestire tutte le idiosincrasie, come l'accordo grammaticale (es. degli articoli). Un lavoro futuro potrebbe definire una convenzione più pratica e flessibile, il che permetterebbe di generare automaticamente varianti dello stesso scenario in maniera economica, aumentando anche la robustezza della valutazione contro possibili bias posizionali o di ruolo.

La seconda direzione riguarda la creazione di nuovi scenari. L'estensione di DiPRomIt-scenarios è un'operazione che richiede un investimento significativo in termini di tempo: la redazione di ogni scenario narrativo è un processo creativo vincolato da linee guida precise (pertinenza culturale, ambiguità controllata, assenza di soluzioni per esclusione). Qualsiasi ampliamento futuro dovrebbe prevedere la partecipazione di più annotatori,

una validazione strutturata del loro lavoro in rispetto delle linee guida di costruzione degli scenari stabilite per DiPRomIt-scenarios, in modo da garantire la coerenza con il dataset esistente.

Il caso di studio del gruppo 372 (Sezione 4.6) suggerisce una direzione di ricerca possibilmente interessante: un’annotazione della struttura metaforica dei proverbi in DiPRomIttra Varianti e Varietà, nei termini della teoria della metafora concettuale (Lakoff & Johnson, 1980). Questo sarebbe possibile individuando, per ciascun proverbio, il dominio sorgente e il dominio target coinvolti e il grado di esplicitezza del *tertium comparationis*. Una tale annotazione permetterebbe di verificare se il grado di esplicitezza della mappatura metaforica sia correlato alla performance dei modelli, e in quale direzione. Goren e Strapparava (2024) mostrano che, nel riconoscimento di metafore a livello di parola nei proverbi inglesi, fornire un contesto aggiuntivo migliora sensibilmente le prestazioni di GPT-3.5 in zero-shot: più informazione contestuale favorisce, in generale, la lettura figurata. DiPRomIt potrebbe costituire un banco di prova per il confronto.

Estensioni del protocollo sperimentale. Sul piano degli esperimenti, lavori futuri potranno esplorare impostazioni few-shot nonché in-context learning³. L’analisi del contributo dei punteggi di confidenza del modello — ovvero della distribuzione di probabilità sulle opzioni A, B, C, D — potrebbe

³ Con *in-context learning* si intende la capacità di un modello di adattare il proprio comportamento sulla base di istruzioni o esempi forniti nel prompt al momento dell’inferenza, senza aggiornarne i parametri: i regimi zero-shot e few-shot ne costituiscono casi particolari, a seconda del numero di esempi inclusi nel prompt (Brown et al., 2020).

inoltre arricchire la valutazione qualitativa degli errori, distinguendo risposte sbagliate legate a incertezza genuina da quelle legate a bias sistematici.

Modelli di ragionamento open-weight. Una direzione interessante riguarda la valutazione degli output di modelli di ragionamento (*reasoning LLM*), ovvero modelli addestrati a produrre una catena di pensiero esplicita (*chain-of-thought*, CoT) prima di emettere la loro risposta finale. A differenza dei modelli standard valutati in questo lavoro, questi modelli operano in modalità multi-step: scompongono il problema, durante questa operazione verificano ipotesi intermedie e revisionano il ragionamento prima di selezionare un’opzione. Tale modalità è teoricamente ben adatta al task di selezione dei proverbi più plausibili presentata qui, tuttavia valutare le risposte e il loro formato potrebbe risultare particolarmente difficoltoso.

Al momento della stesura di questo lavoro, esistono già modelli di ragionamento open-weight maturi e direttamente comparabili con quelli valutati in questo studio, come DeepSeek-R1 (DeepSeek-AI et al., 2025) e QwQ-32B (Qwen Team, 2025b), entrambi disponibili anche in varianti di dimensioni compatibili con il nostro set sperimentale.

L’interesse a questi modelli per il task di DiPRomIt-scenarios è duplice. Da un lato, la catena di pensiero esplicita⁴ (<think>...</think>) consente un’analisi qualitativa degli errori di gran lunga più ricca rispetto alla sola accuracy: è possibile osservare se il modello «riconosce» l’area

⁴ Tuttavia è opportuno precisare che il contenuto del blocco <think> non costituisce un resoconto affidabile del processo di «pensiero»: anche i token generati al suo interno sono predizioni *next-token*. Non sorprende quindi spesso osservare discrepanze tra quanto esplicitato nel blocco <think> e la risposta selezionata dal modello.

geografica, se interpreta correttamente il registro narrativo dello scenario, se discrimina tra i distrattori per ragioni semantiche o per esclusione superficiale. Dall'altro, l'ipotesi teorica che emerge dai nostri risultati — ossia che la difficoltà degli scenari dipenda dall'opacità culturale del proverbio piuttosto che dalla sua distanza geografica — potrebbe essere verificata più direttamente analizzando le tracce di output, probabilmente riconducibile a un «ragionamento».

Più in generale, i risultati di questo lavoro suggeriscono che la sensibilità culturale dei LLM per le varietà linguistiche locali rimane un problema aperto, con ricadute che vanno ben oltre la paremiologia dialettale: la capacità di «comprensione», paragonata a quella umana, rimane fortemente limitata, e generare linguaggio culturalmente radicato è una componente essenziale di qualsiasi sistema NLP che aspiri a operare in contesti linguisticamente plurali.

Appendice

A. Fonti WikiQuote

La Tabella 1 riporta l’elenco completo degli URL WikiQuote da cui sono stati raccolti, curati e arricchiti i proverbi in DiPRomIt, insieme alla regione amministrativa e alle informazioni geografiche di dettaglio incluse nel dataset.

Tabella 1: Elenco degli URL WikiQuote utilizzati per la raccolta di DiPRomIt. **Regione:** regione amministrativa di riferimento per i proverbi dell'URL. **Informazioni geografiche di dettaglio:** province e comuni rappresentati (*Generica:* nessuna informazione geografica di dettaglio disponibile; *: l'informazione geografica corrisponde a un'area che può essere più grande o più piccola di una provincia. In blu: l'informazione geografica in WikiQuote era scritta nel testo del proverbio ed è stata estratta e aggiunta come metadato in DiPRomIt).

Regione	Informazioni geografiche di dettaglio	URL
Abruzzo	<i>Generica</i>	https://it.wikiquote.org/wiki/Proverbi_abruzzesi
Basilicata	<i>Generica</i> , <i>Basilicata meridionale*</i>	https://it.wikiquote.org/wiki/Proverbi_lucani
Calabria	Reggio Calabria	https://it.wikiquote.org/wiki/Proverbi_reggini
Calabria	<i>Generica</i>	https://it.wikiquote.org/wiki/Proverbi_calabresi
Campania	Caggiano (Salerno)	https://it.wikiquote.org/wiki/Proverbi_caggianesi
Campania	Calitri (Avellino)	https://it.wikiquote.org/wiki/Proverbi_calitrani
Campania	Procida (Napoli)	https://it.wikiquote.org/wiki/Proverbi_procidani
Campania	Teggiano (Salerno)	https://it.wikiquote.org/wiki/Proverbi_teggianesi
Campania	<i>Generica</i> , <i>Irpinia*</i> , <i>Salerno</i>	https://it.wikiquote.org/wiki/Proverbi_campani
Campania	Napoli	https://it.wikiquote.org/wiki/Proverbi_napoletani
Emilia-Romagna	Bologna	https://it.wikiquote.org/wiki/Proverbi_bolognesi
Emilia-Romagna	Ferrara	https://it.wikiquote.org/wiki/Proverbi_ferraresi
Emilia-Romagna	Parma	https://it.wikiquote.org/wiki/Proverbi_parmigiani
Emilia-Romagna	Romagna*	https://it.wikiquote.org/wiki/Proverbi_romagnoli
Friuli-Venezia Giulia	<i>Generica</i> , <i>Udine</i> , <i>Socchieve (Udine)</i>	https://it.wikiquote.org/wiki/Proverbi_friulani
Lazio	Gaeta (Latina)	https://it.wikiquote.org/wiki/Proverbi_gaetani
Lazio	Roma	https://it.wikiquote.org/wiki/Proverbi_romani
Liguria	<i>Generica</i>	https://it.wikiquote.org/wiki/Proverbi_liguri
Liguria	Genova	https://it.wikiquote.org/wiki/Proverbi_genovesi
Lombardia	<i>Generica</i>	https://it.wikiquote.org/wiki/Proverbi_lombardi
Lombardia	Bergamo	https://it.wikiquote.org/wiki/Proverbi_bergamaschi
Lombardia	Brescia	https://it.wikiquote.org/wiki/Proverbi_bresciani
Lombardia	Brianza*	https://it.wikiquote.org/wiki/Proverbi_brianzoli
Lombardia	Lodi	https://it.wikiquote.org/wiki/Proverbi_lodigiani
Lombardia	Mantova	https://it.wikiquote.org/wiki/Proverbi_mantovani
Lombardia	Milano	https://it.wikiquote.org/wiki/Proverbi_milanesi
Lombardia	Valtellina*	https://it.wikiquote.org/wiki/Proverbi_valtellinesi
Lombardia	Val Camonica*	https://it.wikiquote.org/wiki/Proverbi_camuni
Marche	<i>Generica</i>	https://it.wikiquote.org/wiki/Proverbi_marchigiani
Molise	<i>Generica</i> , <i>Campobasso</i> , <i>Agnone (Isernia)</i>	https://it.wikiquote.org/wiki/Proverbi_molisani
Piemonte	Monferrato*	https://it.wikiquote.org/wiki/Proverbi_monferrini
Piemonte	<i>Generica</i>	https://it.wikiquote.org/wiki/Proverbi_piemontesi
Puglia	<i>Generica</i>	https://it.wikiquote.org/wiki/Proverbi_pugliesi
Puglia	Lecce	https://it.wikiquote.org/wiki/Proverbi_leccesi
Puglia	Gravina (Bari)	https://it.wikiquote.org/wiki/Proverbi_gravinesi
Puglia	Salento*	https://it.wikiquote.org/wiki/Proverbi_salentini
Puglia	Taranto	https://it.wikiquote.org/wiki/Proverbi_tarantini
Sardegna	<i>Generica</i> , <i>Gallura*</i> , <i>Nuoro</i> , <i>Bosa (Oristano)</i> , <i>Ardara (Sassari)</i> , <i>Bonorva (Sassari)</i> , <i>Nule (Sassari)</i>	https://it.wikiquote.org/wiki/Proverbi_sardi
Sicilia	Palermo	https://it.wikiquote.org/wiki/Proverbi_palermitani
Sicilia	<i>Generica</i> , <i>Erice (Trapani)</i>	https://it.wikiquote.org/wiki/Proverbi_siciliani
Toscana	Livorno	https://it.wikiquote.org/wiki/Proverbi_livornesi
Trentino-Alto Adige	<i>Generica</i>	https://it.wikiquote.org/wiki/Proverbi_trentini
Veneto	<i>Generica</i>	https://it.wikiquote.org/wiki/Proverbi_veneti
Veneto	Venezia	https://it.wikiquote.org/wiki/Proverbi_veneziani

B. Regioni W_{close} per DiPRomIt-scenarios

La Tabella 2 riporta, per ciascuna regione, l'insieme delle regioni da cui viene campionato casualmente un proverbio per costruire l'opzione W_{close} degli scenari in DiPRomIt-scenarios. La selezione è basata su criteri di contiguità geografica e/o affinità linguistica tra varietà dialettali. Umbria e Valle d'Aosta sono assenti da DiPRomIt come anticipatamente detto. La Sardegna rappresenta un caso limite rispetto a entrambi i criteri: non condivide contiguità geografica con nessun'altra regione della tabella, e il sardo, a differenza delle altre varietà qui considerate, costituisce un gruppo linguistico distinto da quello italo-romanzo (Pellegrini, 1977); l'insieme W_{close} per la Sardegna risulta pertanto vuoto.

Tabella 2: Insieme delle regioni da cui viene campionato il proverbio W_{close} per ciascuna regione in DiPRomIt-scenarios.

Regione	Insieme W_{close}
Abruzzo	Marche, Molise, Lazio
Basilicata	Calabria, Campania, Puglia
Calabria	Basilicata, Puglia, Sicilia
Campania	Basilicata, Lazio, Molise, Puglia
Emilia-Romagna	Liguria, Lombardia, Marche, Piemonte, Toscana, Veneto
Friuli-Venezia Giulia	Veneto
Lazio	Abruzzo, Campania, Marche, Molise, Toscana
Liguria	Emilia-Romagna, Piemonte, Toscana
Lombardia	Emilia-Romagna, Piemonte, Trentino-Alto Adige, Veneto
Marche	Abruzzo, Emilia-Romagna, Lazio, Toscana
Molise	Abruzzo, Campania, Lazio, Puglia
Piemonte	Emilia-Romagna, Liguria, Lombardia
Puglia	Basilicata, Calabria, Campania, Molise, Sicilia
Sardegna	—
Sicilia	Calabria, Puglia
Toscana	Emilia-Romagna, Lazio, Marche
Trentino-Alto Adige	Lombardia, Veneto
Veneto	Emilia-Romagna, Friuli-Venezia Giulia, Lombardia, Trentino-Alto Adige

C. Test di Wilcoxon signed-rank: assunzioni e calcolo (Sezione 4.4)

Questa appendice documenta per intero il test di Wilcoxon signed-rank citato nella Sezione 4.4, a scopo di trasparenza e riproducibilità.

Verifica delle assunzioni

Il test richiede tre condizioni⁵:

1. *Variabile dipendente di livello ordinale o continuo*: soddisfatta.

L'accuracy è una proporzione continua nell'intervallo $[0, 1]$.

2. *Due gruppi correlati (misure ripetute sugli stessi soggetti)*: soddisfatta. Ciascuno dei dieci modelli è valutato in entrambe le condizioni geografiche, configurando un disegno a coppie appaiate analogo a una misurazione «prima/dopo».

3. *Distribuzione delle differenze simmetrica attorno alla mediana*: soddisfatta (con qualche precauzione). La skewness delle dieci differenze è $+0,52$ (asimmetria moderata, trainata soprattutto dal guadagno di $Q_{wen-2.5-7B}$, $+0,0285$, il più alto del gruppo); non una violazione grave, ma con $n = 10$ la potenza per un test formale di simmetria è comunque nulla. Per questo motivo, la conclusione è stata verificata anche con un test dei segni (v. nota a fine sezione), che non richiede questa assunzione.

⁵ Cfr. la guida di Laerd Statistics.

Dati e calcolo, passo per passo

Tabella 3: Accuracy nelle due condizioni geografiche per i dieci modelli, differenza Δ e rango assegnato a $|\Delta|$ (nessun pareggio, nessuna differenza nulla).

Modello	without-geo	with-geo	Δ	Rango di $ \Delta $
Mixtral-0.1-8x7B	0.2113	0.2159	+0.0046	1
Llamantino-3-ANITA-8B	0.4272	0.4340	+0.0068	2
Minerva-1.0-7B	0.2730	0.2803	+0.0073	3
Mistral-0.3-7B	0.3206	0.3297	+0.0091	4
Llama-3.1-70B	0.5324	0.5437	+0.0113	5
Gemma-2-27B	0.5206	0.5369	+0.0163	6
Llama-3.1-8B	0.3519	0.3687	+0.0168	7
Qwen-2.5-72B	0.6436	0.6617	+0.0181	8
Gemma-2-9B	0.4485	0.4699	+0.0214	9
Qwen-2.5-7B	0.4572	0.4857	+0.0285	10

La procedura si articola in cinque passaggi:

1. Calcolo della differenza $d_i = \text{acc}_{\text{with-geo},i} - \text{acc}_{\text{without-geo},i}$ per ciascuno dei dieci modelli (colonna Δ in Tabella 3).
2. Si mettono in ordine crescente i valori assoluti $|d_i|$ e si assegna a ciascuno un rango da 1 a 10 (nel nostro caso non sono presenti pareggi).
3. Si ridistribuiscono i segni originali di d_i al rango corrispondente.
4. Si sommano separatamente i ranghi con segno positivo (R^+) e quelli con segno negativo (R^-), con $R^+ + R^- = \frac{n(n+1)}{2} = 55$ per costruzione.
5. Si confronta il calcolo statistico con la sua distribuzione sotto l'ipotesi nulla.

Nel caso in esame, tutti e dieci i segni sono positivi: $R^+ = 55$, $R^- = 0$, lo scarto massimo possibile dalla situazione di parità.

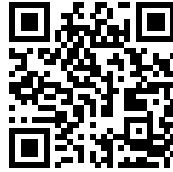
Sotto l'ipotesi nulla, ciascun segno è positivo o negativo con probabilità $1/2$ indipendentemente dagli altri, dando luogo a $2^{10} = 1024$ configurazioni di segno equiprobabili. Il pattern osservato è una sola di queste 1024 configurazioni, ed è la più estrema possibile. Da cui $p = 1/1024 \approx 0,001$ a una coda, o $p = 2/1024 \approx 0,002$ a due code (includendo per simmetria l'ipotetico caso «tutti negativi»).

Il calcolo è stato eseguito con la funzione `scipy.stats.wilcoxon` (Python) sui valori di Tabella 3 (Virtanen et al., 2020); non comportando alcuna componente stocastica, è interamente riproducibile a partire dalla tabella stessa, anche a mano seguendo i cinque passaggi sopra elencati⁶.

⁶ Poiché tutti i segni concordano, un test dei segni che non richiede l'assunzione di simmetria discussa sopra restituisce lo stesso risultato ($p = 2 \times (1/2)^{10} \approx 0,002$), a conferma della correttezza della conclusione.

D. Codice per l'elaborazione dei risultati e la riproducibilità

Il codice per l'elaborazione dei risultati grezzi, il calcolo delle statistiche aggregate e la generazione delle figure discusse nei Capitoli 3 e 4 è reso pubblicamente disponibile in un deposito Zenodo dedicato⁷.



La Tabella 4 riepiloga funzione e collocazione di ciascuno script nella pipeline di analisi, dai risultati grezzi (JSON) alle figure e tabelle discusse nei Capitoli 3 e 4.

Tabella 4: Script di analisi con una breve spiegazione della funzione e i relativi riferimenti nella tesi.

Script	Funzione	Riferimento in tesi
script1_build_matrix.py	Converte i file 'grezzi' dei modelli in JSON, poi in una matrice binaria scenario $\times$ configurazione (risposta corretta/errata).	preelaborazione
script2_aggregate_stats.py	Calcola l'accuracy media aggregata per modello, condizione geografica, seed, tipo di prompt e scenario.	Tabella 3.2
figure_scenari_estremi.py	Genera le figure sui 15 scenari più facili/difficili, la stessa analisi disaggregata per condizione geografica, e la heatmap per singolo modello.	Fig. 3.1, 3.2, 3.3
analisi_bias_posizionale_minerva.py	Verifica la preferenza sistematica di Minerva-1.0-7B per la lettera A tra le opzioni di risposta.	nota su Minerva, Cap. 4
analisi_distractors.py	Estrae, per un sottoinsieme di scenari, la distribuzione delle risposte (corretta/distrattori) per modello e condizione geografica.	caso di studio, §4.6
distractor_summary.py	Riorganizza l'output precedente in tabella pivot per scenario $\times$ condizione.	idem

⁷ <https://doi.org/10.5281/zenodo.21805112>

E. Note tecniche sulla curatela del dataset (§2.1)

Questa appendice documenta due difficoltà tecniche incontrate nell'organizzazione dei dati semi-strutturati raccolti da WikiQuote per DiPRomIt, a scopo di trasparenza e come guida per lavori futuri che affrontino una curatela analoga.

Normalizzazione del separatore |

Il carattere | su WikiQuote indica un'interruzione di riga all'interno dei campi Proverbio e Traduzione. La regola adottata è la seguente: se è già presente una virgola, si rimuove il carattere | e si riporta in minuscolo la lettera iniziale della riga successiva; se la virgola non è presente ma l'a capo la implica, si aggiunge la virgola e, di nuovo, si riporta in minuscolo la lettera iniziale della riga successiva. In entrambi i casi, la lettera maiuscola all'inizio della nuova riga diventa sempre minuscola.

Vanno inseriti in *Note* anche il testo originale di WikiQuote prima di qualsiasi normalizzazione; le corrispondenze in italiano standard; le informazioni extra-linguistiche (es. riferimenti numerici ad apice come [39]); e le segnalazioni di wellerismi, potenziali canti e significati multipli. Tra le informazioni extra-linguistiche rientrano, tra gli altri, i riferimenti numerici ad apice che in WikiQuote rimandano a una nota esplicativa o a una fonte bibliografica: ad esempio, nel proverbio veneziano «*Dòna, gata e bisata*^[5] no mor mai» (ID 386, Dalmedico 1857; *Traduzione*, anche se forse qui si tratta più di una spiegazione: «Stentano a morire»), il rimando [5] accanto

a *bisata* rinvia a una nota WikiQuote che ne chiariva il significato — anguilla; il numero è stato rimosso dal testo del proverbio e l’annotazione spostata nel campo *Note* anche se il rinvio è presente nella fonte originale. Il motivo è stato principalmente di non avere caratteri non alfanumerici nel testo del dataset che potesse essere eventualmente parsato senza intoppi, come avviene spesso nelle tradizioni NLP.

Citazioni indirette

Se WikiQuote riporta un proverbio attestato da un autore che a sua volta lo cita da un’altra opera, si indica quell’autore come fonte primaria nel dataset.

Bibliografia

- Aggarwal, M., Agarwal, A., Goel, A., Hira, M., & Gupta, A. (2026). MORPHOGEN: A Multilingual Benchmark for Evaluating Gender-Aware Morphological Generation.
- AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., & Diab, M. (2024). Investigating Cultural Alignment of Large Language Models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12404–12422. <https://doi.org/10.18653/v1/2024.acl-long.671>
- Alkhamissi, M., Xiao, Y., AlKhamissi, B., & Diab, M. T. (2026). Hire Your Anthropologist! Rethinking Culture Benchmarks Through an Anthropological Lens. In V. Demberg, K. Inui & L. Marquez (Cur.), *Findings of the Association for Computational Linguistics: EACL 2026* (pp. 1218–1235). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2026.findings-eacl.63>
- Almeida, T. S., Bonás, G. K., & Santos, J. G. A. (2025). BRoverbs - Measuring how much LLMs understand Portuguese proverbs. *Journal of the Brazilian Computer Society*, 31(1), 1077–1087. <https://doi.org/10.5753/jbcs.2025.5797>
- Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4), 555–596. <https://doi.org/10.1162/coli.07-034-R2>
- Atreja, S., Ashkinaze, J., Li, L., Mendelsohn, J., & Hemphill, L. (2024). *Prompt Design Matters for Computational Social Science Tasks but in Unpredictable Ways* (1). <https://doi.org/10.48550/ARXIV.2406.11980>

- Azime, I. A., Tonja, A. L., Belay, T. D., Chanie, Y., Balcha, B. F., Abadi, N. H., Ademtew, H. B., Nerea, M. A., Yadeta, D. D., Geremew, D. D., Tesfu, A. A., Slusallek, P., Solorio, T., & Klakow, D. (2025). ProverbEval: Exploring LLM Evaluation Challenges for Low-resource Language Understanding. *Findings of the Association for Computational Linguistics: NAACL 2025*, 6250–6266. <https://doi.org/10.18653/v1/2025.findings-naacl.350>
- Bartoli, M., Pellis, U., & Massobrio, L. (1995). *Atlante Linguistico Italiano*. Istituto Poligrafico e Zecca dello Stato.
- Beccaria, G. L. (Cur.). (2004). *Dizionario di linguistica e di filologia, metrica, retorica*. Einaudi.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*, 610–623.
- Berruto, G. (2014). Esiste ancora l'italiano popolare? Una rivisitazione. In P. Danler & C. Konecny (Cur.), *Dall'architettura della lingua italiana all'architettura linguistica dell'Italia. Saggi in omaggio a Heidi Siller-Runggaldier* (pp. 277–290). Peter Lang.
- Berruto, G., & Cerruti, M. (2015). *La linguistica. Un corso introduttivo* (2^a ed.). UTET Università.
- Berti, C. C. (1874). *Vocabolario bolognese italiano*. Stabilimento tipografico di G. Monti.
- Boccaccio, G. (1980). *Decameron* (V. Branca, Cur.) [Edizione critica]. Einaudi.
- Borenstein, N., Arora, A., Kaffee, L.-A., & Augenstein, I. (2025). Investigating Human Values in Online Communities. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. <https://doi.org/10.18653/v1/2025.naacl-long.77>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language Models Are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*. <https://doi.org/10.48550/arXiv.2005.14165>

- Bulté, B., & Rigouts Terryn, A. (2026). LLMs and Cultural Values: The Impact of Prompt Language and Explicit Cultural Framing. *Computational Linguistics*. <https://doi.org/10.1162/COLI.a.583>
- Calderaro, S., & Monti, J. (2026). A Bolu: A Structured Dataset for the Computational Analysis of Sardinian Improvisational Poetry. In A. Anastasopoulos, S. Markantonatou, A. Ralli, M. Zampieri, S. Bompolas & V. Stamou (Cur.), *Proceedings of the First Workshop on Dialects in NLP — A Resource Perspective* (pp. 1–11). European Language Resources Association (ELRA). <https://doi.org/10.63317/3d9ufzotw55s>
- Cappello, C. M., D’Ali, S., Guglielmetti, M., Di Nuovo, E., & Bosco, C. (2025). Arbuli Sunnu: A Sicilian-Italian Parallel Treebank. In C. Bosco, E. Jezek, M. Polignano & M. Sanguinetti (Cur.), *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)* (pp. 155–168). CEUR Workshop Proceedings. <https://aclanthology.org/2025.clicit-1.17/>
- Cerruti, M., & Regis, R. (2005). Code Switching e Teoria Linguistica: La Situazione Italo-Romanza. *Italian Journal of Linguistics*, 17(1), 179–208.
- Cristianini, N. (2023). *La scorciatoia. Come le macchine sono diventate intelligenti senza pensare in modo umano*. il Mulino.
- Culicover, P. W., Jackendoff, R., & Audring, J. (2017). Multiword Constructions in the Grammar. *Topics in cognitive science*, 9 3(3), 552–568. <https://doi.org/10.1111/tops.12255>
- DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., et al. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. <https://arxiv.org/abs/2501.12948>
- Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). *LLM.Int8(): 8-Bit Matrix Multiplication for Transformers at Scale* (2). <https://doi.org/10.48550/ARXIV.2208.07339>
- Dettmers, T., & Zettlemoyer, L. (2022). *The Case for 4-Bit Precision: K-Bit Inference Scaling Laws* (2). <https://doi.org/10.48550/ARXIV.2212.09720>

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT. Pre-training of Deep Bidirectional Transformers for Language Understanding. In J. et al. Burstein (Cur.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 4171–4186). Association for Computational Linguistics.
- Dickins, J. (2020). An Ontology for Collocations, Formulaic Sequences, Multiword Expressions, Compounds, Phrasal Verbs, Idioms and Proverbs. *Linguistica Online*, 23, 29–72.
- Dong, G., Sun, J., Wang, H., & Wang, X. (2024). *Evaluating and Mitigating Linguistic Discrimination in Large Language Models*. arXiv: 2404.18534 [cs.CL]. Recuperato luglio 30, 2026, da <https://arxiv.org/abs/2404.18534>
- Dror, R., Peled-Cohen, L., Shlomov, S., & Reichart, R. (2020). *Statistical Significance Testing for Natural Language Processing*. Springer International Publishing. <https://doi.org/10.1007/978-3-031-02174-9>
- Errico, G., Ramponi, A., Tonelli, S., & Ježek, E. (2026). Can LLMs Match Folk Wisdom? A Contextual Evaluation of Proverb Selection in Romance Dialects of Italy. *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2026)*.
- Faisal, F., Ahia, O., Srivastava, A., Ahuja, K., Chiang, D., Tsvetkov, Y., & Anastasopoulos, A. (2024). DIALECTBENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14412–14454. <https://doi.org/10.18653/v1/2024.acl-long.777>
- Fillmore, C. J., & Baker, C. (2012). A Frames Approach to Semantic Analysis. In B. Heine & H. Narrog (Cur.), *The Oxford Handbook of Linguistic Analysis* (1^a ed., pp. 313–340). Oxford University Press. <https://doi.org/10.1093/oxfordhob/9780199544004.013.0013>
- Fiorentini, I. (2022). *Sociolinguistica delle minoranze in Italia: un'introduzione* (1a edizione). Carocci editore.
- Franceschi, T. (1994). Il proverbio e la Scuola Geoparemiologica Italiana. *Paremia*, 3, 27–36.

- Franceschi, T. (2004). La formula proverbiale. In V. Boggione & L. Massobrio (Cur.), *Dizionario dei proverbi. I proverbi italiani organizzati per temi* (pp. IX–XXVI). UTET.
- Franceschi, T. (2007). L’Atlante Paremiologico Italiano. In N. Maraschio (Cur.), *Firenze e la lingua italiana fra nazione ed Europa*. Firenze University Press.
- Franceschi, T., Cervini, C., Guarino Amato, M., Mazo, M., & Melis, L. (2000). *Atlante Paremiologico Italiano: Questionario. Ventimila detti proverbiali raccolti in ogni regione d’Italia* (2^a ed.). Edizioni dell’Orso.
- Frary, R. B. (1988). Formula Scoring of Multiple-Choice Tests (Correction for Guessing). *Educational Measurement: Issues and Practice*, 7(2), 33–38. <https://doi.org/10.1111/j.1745-3992.1988.tb00434.x>
- Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100, 25–50.
- Gemma Team, Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. (2024). Gemma 2: Improving Open Language Models at a Practical Size. <https://arxiv.org/abs/2408.00118>
- Ghosh, S., & Srivastava, S. (2022). ePiC: Employing Proverbs in Context as a Benchmark for Abstract Language Understanding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3990–4003. <https://doi.org/10.18653/v1/2022.acl-long.276>
- Giacomelli, G. (2000). *Atlante Lessicale Toscano (versione in cd-rom)*. Lexis Progetti Editoriali.
- Giora, R. (1997). Understanding figurative and literal language: The graded salience hypothesis. *cogl*, 8(3), 183–206. <https://doi.org/10.1515/cogl.1997.8.3.183>
- Goren, G., & Strapparava, C. (2024). Context Matters: Enhancing Metaphor Recognition in Proverbs. *Proceedings of the Language Resources and*

- Evaluation Conference*, 3825–3830. <https://doi.org/10.63317/2abvgi94k5ih>
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Laskar, A., Tsimpoukelli, M., et al. (2024). The Llama 3 Herd of Models. <https://arxiv.org/abs/2407.21783>
- Guo, S., Jiang, S., He, Q., Xiao, Y., Liang, J., Yude, B., He, M., Tao, S., & Zhang, L. (2025). Do Large Language Models Truly Understand Cross-cultural Differences? <https://arxiv.org/abs/2512.07075>
- Haberland, C. R., Maillard, J., & Lusito, S. (2024). Italian-Ligurian Machine Translation in Its Cultural Context. In M. Melero, S. Sakti & C. Soria (Cur.), *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024* (pp. 168–176). ELRA; ICCL. <https://aclanthology.org/2024.sigul-1.21/>
- He, W., Vieira, T. K., Garcia, M., Scarton, C., Idiart, M., & Villavicencio, A. (2025). Investigating Idiomaticity in Word Representations. *ArXiv*, *abs/2411.02610*. <https://doi.org/10.48550/arxiv.2411.02610>
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring Massive Multitask Language Understanding. *International Conference on Learning Representations (ICLR)*. <https://doi.org/10.48550/arXiv.2009.03300>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The Weirdest People in the World? *Behavioral and Brain Sciences*, 33(2–3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hosseini, F., Yousefzadeh, M., & Yaghoobzadeh, Y. (2026). FFE-Hallu: Hallucinations in Fixed Figurative Expressions: A Benchmark of Idioms and Proverbs in the Persian Language. *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5222–5235. <https://doi.org/10.18653/v1/2026.eacl-long.241>
- Hymes, D. (1974). *Foundations in Sociolinguistics. An Ethnographic Approach*. Tavistock.

- Jaberg, K., Jud, J., & Sanga, G. (1987). *Atlante Linguistico ed Etnografico dell'Italia e della Svizzera Meridionale* (Italian). Unicopli.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). Mistral 7B. <https://arxiv.org/abs/2310.06825>
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., ... El Sayed, W. (2024). Mixtral of Experts. <https://arxiv.org/abs/2401.04088>
- Kaasa, A., & Welzel, C. (2023). Elements of Schwartz’s Model in the WVS: How Do They Relate to Other Cultural Models? *Cross-Cultural Research*, 57(5), 431–471. <https://doi.org/10.1177/10693971231179792>
- Kabir, M., Abrar, A., & Ananiadou, S. (2025). Break the Checkbox: Challenging Closed-Style Evaluations of Cultural Alignment in LLMs. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 24–51. <https://doi.org/10.18653/v1/2025.emnlp-main.2>
- Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large Language Models Struggle to Learn Long-Tail Knowledge. *Proceedings of the 40th International Conference on Machine Learning*, 202, 15696–15707.
- Katz, A. N., & Ferretti, T. R. (2001). Moment-By-Moment Reading of Proverbs in Literal and Nonliteral Contexts. *Metaphor and Symbol*, 16(3–4), 193–221. <https://doi.org/10.1080/10926488.2001.9678895>
- Kloss, H. (1978). *Die Entwicklung Neuer Germanischer Kultursprachen Seit 1800*. Schwann.
- Kommers, C., Ahnert, R., Antoniak, M., Benetos, E., Benford, S., Bunz, M., Caramiaux, B., Concannon, S., Disley, M., Dobson, J., Du, Y., Duéñez-Guzmán, E., Francksen, K., Gius, E., Gray, J., Heuser, R., Immel, S., So, R., Leigh, S., ... Hemment, D. (2025). *Computational Hermeneutics*:

- Evaluating Generative AI as a Cultural Technology*. <https://doi.org/10.2139/ssrn.5409144>
- Kurath, H. (1949). *A word geography of the eastern united states*. University of Michigan Press.
- Lakoff, G., & Johnson, M. (1980). *Metaphors We Live By*. University of Chicago Press.
- Lemghari, E. M. (2021). Constructing a Broad Model for Proverb Understanding. *Metaphor and Symbol*, 36(4), 265–287. <https://doi.org/10.1080/10926488.2021.1933489>
- Liu, C. C., Gurevych, I., & Korhonen, A. (2025). Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art. *Transactions of the Association for Computational Linguistics*, 13, 652–689. https://doi.org/10.1162/tacl_a_00760
- Liu, C. C., Koto, F., Baldwin, T., & Gurevych, I. (2024). Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings. *ArXiv, abs/2309.08591*. <https://doi.org/10.48550/arxiv.2309.08591>
- Lu, Y., Keung, P., Ladhak, F., Bhardwaj, V., Zhang, S., & Sun, J. (2018). A neural interlingua for multilingual machine translation. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 84–92. <https://doi.org/10.18653/v1/W18-6309>
- Lutz, M., Sen, I., Ahnert, G., Rogers, E., & Strohmaier, M. (2025). The Prompt Makes the Person(a): A Systematic Evaluation of Sociodemographic Persona Prompting for Large Language Models. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 23212–23237. <https://doi.org/10.18653/v1/2025.findings-emnlp.1261>
- Madaan, L., Singh, A. K., Schaeffer, R., Poulton, A., Koyejo, S., Stenetorp, P., Narang, S., & Hupkes, D. (2024). Quantifying Variance in Evaluation Benchmarks. *arXiv preprint arXiv:2406.10229*. <https://doi.org/10.48550/arXiv.2406.10229>

- Magdy, S. M., Kwon, S. Y., Alwajih, F., Abdelfadil, S. T., Shehata, S., & Abdul-Mageed, M. (2025). JAWAHER: A Multidialectal Dataset of Arabic Proverbs for LLM Benchmarking. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 12320–12341. <https://doi.org/10.18653/v1/2025.naacl-long.613>
- Masoud, R. I., Liu, Z., Ferianc, M., Treleaven, P., & Rodrigues, M. (2025). Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede’s Cultural Dimensions. *Proceedings of the 31st International Conference on Computational Linguistics*, 8474–8503. <https://aclanthology.org/2025.coling-main.567/>
- McEnery, T., & Xiao, R. (2007). Chapter 2. Parallel and Comparable Corpora: What Is Happening? In G. Anderman & M. Rogers (Cur.), *Incorporating Corpora* (pp. 18–31). Multilingual Matters. <https://doi.org/10.21832/9781853599873-005>
- Mensa, E., Zane, L., Scozzaro, C. J., Delsanto, M., Milani, T., & Radicioni, D. P. (2026). ProverbIT - Easy to complete, hard to choose: A CALAMITA Challenge. *Proceedings of the 9th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026)*. <https://ceur-ws.org/Vol-4195/70.pdf>
- Mensa, E., Zane, L., Scozzaro, C. J., Delsanto, M., Milani, T., & Radicioni, D. P. (2025). Easy to Complete, Hard to Choose: Investigating LLM Performance on the ProverbIT Benchmark. *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, 722–734.
- Mieder, W. (2004). *Proverbs: A handbook* (1. publ). Greenwood.
- Mitkov, R. (Cur.). (2014, aprile 1). *The Oxford Handbook of Computational Linguistics 2nd edition* (2^a ed.). Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199573691.001.0001>
- Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., & Stanovsky, G. (2024). State of What Art? A Call for Multi-Prompt LLM Evaluation.

- Transactions of the Association for Computational Linguistics*, 12, 933–949. <https://aclanthology.org/2024.tacl-1.52>
- Myrzakhan, A., Bsharat, S. M., & Shen, Z. (2024). Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena. *arXiv preprint arXiv:2406.07545*. <https://doi.org/10.48550/arXiv.2406.07545>
- Oh, J., Cha, I., Saxon, M., Lim, H., Bhatt, S., & Oh, A. (2025). *Culture Is Everywhere: A Call for Intentionally Cultural Evaluation* (2). <https://doi.org/10.48550/ARXIV.2509.01301>
- Orlando, R., Huguet Cabot, P.-L., & Navigli, R. (2024). Minerva: A LLM-Based Agent for Navigating Complex Scenarios Across Diverse Domains [Identificatore HF: sapienzanlp/Minerva-7B-instruct-v1.0]. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. <https://arxiv.org/abs/2410.00200>
- Özbal, G., Strapparava, C., & Tekiroglu, S. S. (2016). PROMETHEUS: A Corpus of Proverbs Annotated with Metaphors. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.
- Partee, B. H. (1995). Lexical Semantics and Compositionality. In L. R. Gleitman & M. Liberman (Cur.), *An Invitation to Cognitive Science: Language* (pp. 311–360). MIT Press.
- Pavlopoulos, J., Louridas, P., & Filos, P. (2024). Towards a Greek Proverb Atlas: Computational Spatial Exploration and Attribution of Greek Proverbs. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11842–11854. <https://doi.org/10.18653/v1/2024.emnlp-main.661>
- Pawar, S., Park, J., Jin, J., Arora, A., Myung, J., Yadav, S., Haznitrama, F. G., Song, I., Oh, A., & Augenstein, I. (2025a). Survey of Cultural Awareness in Language Models: Text and Beyond. *Computational Linguistics*, 51(3), 907–1004. <https://doi.org/10.1162/COLI.a.14>

- Pawar, S., Park, J., Jin, J., Arora, A., Myung, J., Yadav, S., Haznitrana, F. G., Song, I., Oh, A., & Augenstein, I. (2025b). Survey of Cultural Awareness in Language Models: Text and Beyond. *Computational Linguistics*, 51(3), 907–1004. <https://doi.org/10.1162/coli.a.14>
- Pellegrini, G. B. (1977). *Carta dei Dialetti d'Italia*. Pacini.
- Pezeshkpour, P., & Hruschka, E. (2024). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In K. Duh, H. Gomez & S. Bethard (Cur.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 2006–2017). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-naacl.130>
- Pimpalgaonkar, S., Lele, D., Kulkarni, M., & Bhattacharyya, P. (2021). Introduction to ProverbNet: An Online Multilingual Database of Proverbs and Comprehensive Metadata. *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, 638–650. <https://aclanthology.org/2021.icon-main.78>
- Pitrè, G. (1880). *Proverbi siciliani: raccolti e confrontati con quelli degli altri dialetti d'Italia*. Luigi Pedone Lauriel Editore.
- Polignano, M., Basile, P., de Gemmis, M., & Semeraro, G. (2024). LLaMAntino-3-ANITA: An Advanced Instruction-Tuned Italian LLaMA-3 LLM. *Proceedings of the Sixth Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. <https://arxiv.org/abs/2405.07617>
- Pustejovsky, J. (2002). *The generative lexicon* (1. MIT Press paperback ed., 4. print). MIT Press.
- Qiu, H., Fabbri, A. R., Agarwal, D., Huang, K.-H., Tan, S., Peng, N., & Wu, C.-S. (2025). Evaluating Cultural and Social Awareness of LLM Web Agents. *Findings of the Association for Computational Linguistics: NAACL 2025*, 3978–4005. <https://doi.org/10.18653/v1/2025.findings-naacl.222>
- Qwen Team. (2025a). Qwen2.5 Technical Report. <https://arxiv.org/abs/2412.15115>

- Ramponi, A. (2024). Language Varieties of Italy: Technology Challenges and Opportunities. *Transactions of the Association for Computational Linguistics*, 12, 19–38. https://doi.org/10.1162/tacl_a_00631
- Ramponi, A., & Casula, C. (2023a). DiatopIt: A Corpus of Social Media Posts for the Study of Diatopic Language Variation in Italy. In Y. Scherrer, T. Jauhainen, N. Ljubešić, P. Nakov, J. Tiedemann & M. Zampieri (Cur.), *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)* (pp. 187–199). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.vardial-1.19>
- Ramponi, A., & Casula, C. (2023b). GeoLingIt at EVALITA 2023: Overview of the Geolocation of Linguistic Variation in Italy Task. In M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli & G. Venturi (Cur.), *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*. CEUR.org. <https://ceur-ws.org/Vol-3473/paper12.pdf>
- Read, J., Dridan, R., Oepen, S., & Solberg, L. J. (2012). Sentence Boundary Detection: A Long Solved Problem? *Proceedings of COLING 2012: Posters*, 985–994. Recuperato agosto 19, 2026, da <https://aclanthology.org/C12-2096/>
- Schwartz, S. H. (1992). Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In M. P. Zanna (Cur.), *Advances in Experimental Social Psychology* (pp. 1–65, Vol. 25). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60281-6](https://doi.org/10.1016/S0065-2601(08)60281-6)
- Schwartz, S. H. (2012). An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1). <https://doi.org/10.9707/2307-0919.1116>
- Shanahan, M. (2024). Talking about Large Language Models. *Communications of the ACM*, 67(2), 68–79. <https://doi.org/10.1145/3624724>
- Smith, A. L., Greaves, F., & Panch, T. (2023). Hallucination or Confabulation? Neuroanatomy as metaphor in Large Language Models (S. Berkovsky,

- Cur.). *PLOS Digital Health*, 2(11), e0000388. <https://doi.org/10.1371/journal.pdig.0000388>
- Socolof, M. (2024). Partial Compositionality. *Unknown journal*.
- Sprugnoli, R., Moretti, G., Muscianisi, D. G., & Litta, E. (2025). Ciallabacialla! Modeling and Linking a Regional Lexical Resource to Include Sicilian in the Semantic Web. In C. Bosco, E. Jezek, M. Polignano & M. Sanguinetti (Cur.), *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)* (pp. 1093–1101). CEUR Workshop Proceedings. <https://aclanthology.org/2025.clicit-1.103/>
- Tang, X., Duan, X., & Cai, Z. G. (2026). Do Large Language Models Plan Answer Positions? Position Bias in Multiple-Choice Question Generation. *arXiv preprint arXiv:2605.01846*. <https://doi.org/10.48550/arXiv.2605.01846>
- Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural Bias and Cultural Alignment of Large Language Models. *PNAS Nexus*, 3(9), pgae346. <https://doi.org/10.1093/pnasnexus/pgae346>
- Temple, J. G., & Honeck, R. P. (1999). Proverb Comprehension: The Primacy of Literal Meaning. *Journal of Psycholinguistic Research*, 28(1), 41–70. <https://doi.org/10.1023/A:1023287420088>
- Thapa, S., Rauniyar, K., Veeramani, H., Adhikari, S., Razzak, I., & Naseem, U. (2025). Probing the Limits of Multilingual Language Understanding: Low-Resource Language Proverbs as LLM Benchmark for AI Wisdom. *Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025)*, 120–129. <https://doi.org/10.18653/v1/2025.codi-1.11>
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>

- Vo, T., et al. (2025). CURE: Cultural Understanding and Reasoning Evaluation — A Framework for “Thick” Culture Alignment Evaluation in LLMs. <https://arxiv.org/abs/2511.12014>
- Wang, M., Pham, V., Moghimifar, F., & Vu, T.-T. (2025). Proverbs Run in Pairs: Evaluating Proverb Translation Capability of Large Language Model. *ArXiv*, 1646–1662. <https://doi.org/10.48550/arxiv.2501.11953>
- Wang, M., Adel, H., Lange, L., Liu, Y., Nie, E., Strötgen, J., & Schuetze, H. (2025). Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5075–5094. <https://doi.org/10.18653/v1/2025.acl-long.253>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. <https://aclanthology.org/2020.emnlp-demos.6>
- Xu, S., Leng, Y., Yu, L., & Xiong, D. (2025). Self-Pluralising Culture Alignment for Large Language Models. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. <https://doi.org/10.18653/v1/2025.naacl-long.350>
- Yao, J., Yi, X., Gong, Y., Wang, X., & Xie, X. (2024). Value FULCRA: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Value. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8762–8785. <https://doi.org/10.18653/v1/2024.naacl-long.486>

- Zhang, W., Cai, H., & Chen, W. (2025). Beyond the Singular: Revealing the Value of Multiple Generations in Benchmark Evaluation. *arXiv preprint arXiv:2502.08943*. <https://doi.org/10.48550/arXiv.2502.08943>
- Zheng, C., Zhou, H., Meng, F., Zhou, J., & Huang, M. (2024). Large Language Models Are Not Robust Multiple Choice Selectors. *The Twelfth International Conference on Learning Representations*. <https://doi.org/10.48550/arXiv.2309.03882>
- Zhou, N., Bamman, D., & Bleaman, I. L. (2025). *Culture Is Not Trivia: Sociocultural Theory for Cultural NLP* (2). <https://doi.org/10.48550/ARXIV.2502.12057>

Sitografia

- Errico, G. (2026). Correlazioni_Proverbi [Data set]. <https://doi.org/10.5281/zenodo.20444846>
- Qwen Team. (2025b). QwQ-32B: Embracing the Power of Reinforcement Learning [Technical blog post, Alibaba Cloud]. <https://qwenlm.github.io/blog/qwq-32b/>
- World Values Survey Association. (2025). World Values Survey. Recuperato maggio 27, 2025, da <https://www.worldvaluessurvey.org/>

Il primo ringraziamento va alla mia famiglia, e in particolare ai miei genitori, che in questi anni mi hanno sopportato, nel senso più letterale del termine. Un ringraziamento sentito va alla mia relatrice, la Prof.ssa Elisabetta Ježek, per avermi accompagnato e ispirato lungo il percorso. Grazie alla mia ragazza, che mi ha sopportato e supportato per tutto questo tempo: le ho parlato di questa tesi così tante volte che credo possa discuterla al posto mio. Un grazie speciale va a FBK, che ha reso possibile questo lavoro, e in particolare alla Dott.ssa Sara Tonelli e al Dott. Alan Ramponi, che mi hanno seguito passo dopo passo in ogni fase della ricerca, ogni volta che è stato possibile. Un grazie va anche al corpo docenti, che spesso mi ha messo nella condizione di fermarmi a rivalutare i miei interessi e la mia situazione. Associa i loro insegnamenti all'espressione:

To stop and smell the roses.

Infine, un grazie a tutti coloro, diretti e indiretti, che hanno reso possibile questo lavoro e il percorso che lo ha preceduto.