Confronto tra diversi tipi di valutazione del miglioramento della chiarezza di testi amministrativi in lingua italiana

Mariachiara Pascucci[1,†], Mirko Tavosanis[1,*,†]

[1] *Università di Pisa, Dipartimento di Filologia, letteratura e linguistica*

Abstract

The paper presents a comparison of different types of evaluation of administrative texts in the Italian language on which a clarity improvement intervention was carried out. The clarity improvement was performed by human experts and ChatGPT. The evaluation was carried out in four different ways: by expert evaluators, used as a reference; by evaluators with good skills, subject to dedicated training; by generic evaluators recruited through a crowdsourcing platform; by ChatGPT. The results show that the closest match to the results of the evaluation by expert evaluators was reached, by a wide margin, by evaluators with good skills and dedicated training; the second best approach was reached by requesting evaluation from ChatGPT; the worst approach was reached by generic evaluators recruited through a crowdsourcing platform. Task features that may have influenced the outcome are also discussed.

Keywords

Text simplification, LLMs, ChatGPT, Italian, evaluation, crowdsourcing

1. Introduzione

La diffusione dei sistemi di intelligenza artificiale generativa ha portato a una grande richiesta di valutazione delle loro capacità. Il tipo di valutazione universalmente considerato più valido rimane in generale quello realizzato da esseri umani, che però in pratica può essere condotto in modi diversi e con risultati di valore molto diverso. Per alcune capacità, inoltre, non esistono ancora quadri di valutazione condivisi. Rientra senz'altro in quest'ultima categoria anche la valutazione del miglioramento complessivo della chiarezza dei testi in lingua italiana, oggetto dell'analisi qui descritta. Gli indici oggettivi esistenti per l'analisi di testi, come il GULPEASE o la quantificazione delle parole che rientrano nel Vocabolario di Base, descrivono in effetti solo aspetti limitati di un qualunque testo. Per la chiarezza in sé, mentre abbondano le indicazioni su come scrivere in modo chiaro (una sintesi aggiornata è esposta in [1]), non sono mai stati codificati criteri di ampio consenso per la valutazione dei prodotti [2].

Naturalmente, molti metodi di valutazione attuali forniscono almeno un primo orientamento nella maggior parte dei casi. Per esempio, [3] ha mostrato che attraverso il crowdsourcing è possibile ottenere un'indicazione generica ma attendibile sul miglioramento della chiarezza di testi in lingua inglese. Tuttavia, gli studi sull'efficacia di simili pratiche sono ancora poco numerosi ed è senz'altro molto sentita la necessità di migliorare il livello attuale delle conoscenze.

Il presente contributo si inserisce in questo contesto in quanto mette a confronto diversi metodi per valutare il miglioramento della chiarezza dei testi. Oggetto della valutazione sono stati testi piuttosto ampi, rappresentativi dell'italiano amministrativo e resi più chiari attraverso un intervento umano e attraverso la riformulazione con ChatGPT (versione 3.5); il contesto, che ha visto la realizzazione di diverse attività di valutazione collegate, è descritto in dettaglio in [4].

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

* Corresponding author.

† Il contributo degli autori è unitario. Tuttavia, si dichiara che sono opera di Mariachiara Pascucci i paragrafi 2, 3, 4 e 7 e di Mirko Tavosanis i paragrafi 1, 5, 6 e 8.

mariachiara.pascucci@phd.unipi.it (M. Pascucci); mirko.tavosanis@unipi.it (M. Tavosanis).
0009-0007-1934-8479 (M. Pascucci); 0000-0002-4730-3901 (M. Tavosanis)

Ai fini del presente contributo, la valutazione è stata condotta in quattro modi diversi: da valutatori esperti, usati come riferimento; da valutatori con buone competenze, oggetto di una formazione dedicata; da valutatori generici reclutati attraverso una piattaforma di crowdsourcing; da parte di ChatGPT. In tutti i casi, è stata usata la stessa serie di indicazioni per la valutazione. I risultati sono stati analizzati in [4] per le informazioni che forniscono riguardo alla capacità di sistemi come ChatGPT di migliorare efficacemente la chiarezza dei testi. In questa sede si mostrerà invece, in modo più specifico, la differenza nei giudizi in rapporto ai quattro modi di valutazione.

2. Lavori correlati

Anche se il miglioramento della chiarezza è un obiettivo centrale in vari campi di ricerca linguistica applicata, la valutazione dell'efficacia dei processi di miglioramento rimane, come si è detto, una questione aperta. Tale stato di cose si riflette nell'eterogeneità delle soluzioni adottate nei diversi studi realizzati in questo ambito. Come evidenziato in [2], infatti, non esiste un quadro teorico condiviso per valutare l'efficacia delle riformulazioni in termini di chiarezza né, in senso più ampio, per la valutazione complessiva della qualità dei testi generati. In una rassegna sistematica, [5] sottolinea che le operazioni di valutazione dei testi generati possono avvalersi di diversi approcci: valutazioni umane, metriche quantitative o sistemi di valutazione automatica e semiautomatica. Il giudizio umano è adoperato, per esempio, in lavori come [6] e in studi che hanno adottato un approccio comparativo, come [7], che propone un confronto tra valutazione umana e metriche automatiche per valutare l'efficacia dei processi di semplificazione. La letteratura di riferimento sembra in effetti convergere verso l'idea che la valutazione umana dei testi generati rimanga in generale la più adeguata, come evidenziato da diversi lavori, tra cui [8] e [2]. Non mancano tuttavia studi che usano metriche automatiche e indici di leggibilità per la valutazione degli output, come [9].

Riguardo all'impiego del crowdsourcing, un approccio interessante è quello di lavori come il già citato [3] e il più recente [10], in cui sono messe a confronto diverse modalità di valutazione, incluse metriche automatiche, giudizi di esperti e un test di comprensione che ha coinvolto partecipanti selezionati in modo casuale e senza preparazione specifica per lo svolgimento del compito. In ambito italiano, [11] ha esplorato l'uso del crowdsourcing per la valutazione della complessità frasale.

L'applicazione dei modelli GPT alla valutazione automatica della chiarezza testuale è stata ancora poco indagata, ma non mancano gli esperimenti interessanti. Degno di nota è il già citato [7], che ha esaminato il potenziale di GPT-4, confrontando i risultati delle valutazioni del modello con quelle di esperti umani per i processi di semplificazione.

3. Testi originali e riformulazioni

La valutazione cui si fa riferimento nel presente contributo è stata eseguita in rapporto a un'attività di miglioramento della chiarezza di testi amministrativi regolativi in lingua italiana. Questo tipo di attività corrisponde a una richiesta diffusa a livello sociale e su cui esiste ampia bibliografia specifica (per esempio: [12]). Tuttavia, anche in questo caso non esistono criteri condivisi per la valutazione di testi esistenti; non è quindi possibile, per esempio, rifarsi a scale condivise per descrivere la chiarezza di un testo amministrativo. Sulla situazione generale dei criteri per la chiarezza e sui dettagli del caso esaminato si rimanda di nuovo a [1] e [4]; le informazioni fornite qui di seguito saranno quindi solo quelle strettamente necessarie per l'inquadramento dell'esperienza svolta.

Per l'attività descritta qui di seguito sono state scelte casualmente 8 sezioni ragionevolmente autonome e autoconsistenti di testi amministrativi più ampi, per una lunghezza approssimativa di 2000 caratteri a sezione. I testi sono stati poi rielaborati chiedendo a ChatGPT di migliorarne la chiarezza. I due prompt usati per la versione definitiva del lavoro sono riportati nell'Appendice A.

In aggiunta al miglioramento della chiarezza da parte ChatGPT, uno degli autori (Mariachiara Pascucci) ha condotto un intervento umano, usato come termine di confronto, per il miglioramento della chiarezza. Inoltre, nel campione sono stati inseriti, con minimi ritocchi, alcuni esempi classici di miglioramento della chiarezza, ripresi da [13].

4. Interventi di riformulazione

Per quanto riguarda la riformulazione manuale dei testi, gli interventi hanno interessato vari tratti linguistici (a livello lessicale, morfosintattico e testuale) comunemente associati alla complessità dei testi istituzionali. Il quadro di riferimento è quello presentato in [14].

L'analisi delle riformulazioni generate da ChatGPT mostra che il modello ha operato in modo paragonabile a quello umano, intervenendo contemporaneamente su più tratti e su più livelli linguistici. ChatGPT sembra comunque essersi concentrato sulla semplificazione del lessico, spesso piuttosto spinta, e sulla riduzione della lunghezza delle frasi. Di seguito, si riporta un esempio che consente di confrontare la versione originale con i due diversi tipi di riformulazione.

Originale (ENERG-2)

Le spese per “Servizi esterni” sono rappresentate dalle spese che il Beneficiario/Soggetto Attuatore sostiene a favore di erogatori esterni di servizi, i quali si assumono determinati compiti che sono necessari per il raggiungimento degli obiettivi progettuali e che il Beneficiario/Soggetto Attuatore non è in grado di svolgere in proprio.

Riformulazione manuale

Le spese per “Servizi esterni” sono le spese che il Beneficiario/Soggetto Attuatore sostiene a favore di erogatori esterni di servizi. Tali erogatori svolgono compiti specifici, necessari per il raggiungimento degli obiettivi del progetto, che il Beneficiario/Soggetto Attuatore non è in grado di svolgere in proprio.

Riformulazione automatica (Prompt 2)

Le spese per 'Servizi esterni' sono i soldi che una persona o un'organizzazione spende per ottenere aiuto da altri fornitori di servizi. Questi fornitori svolgono compiti importanti per raggiungere gli obiettivi di un progetto, compiti che la persona o l'organizzazione che riceve l'aiuto non può fare da sola.
(Generated by AI tool ChatGPT-3.5)

5. Griglia di valutazione e valutatori

Il primo passo per l’attività è stata la creazione di una griglia di valutazione basata sulla bibliografia esistente e sull’esame diretto delle capacità del sistema. La griglia è stata messa a punto attraverso una serie di verifiche intermedie ed è stata corredata da istruzioni applicative ricavate dalla pratica, con discussione di esempi specifici e indicazioni per la gestione di casi dubbi. La versione definitiva della griglia e delle istruzioni, usata per tutte le attività descritte qui di seguito, è riportata nell’Appendice B.

6. Modalità della valutazione

La valutazione è stata condotta in quattro modi diversi, presentati qui di seguito.

Il quadro concettuale usato è quello descritto in [15]. Come punto di riferimento sono quindi stati usati i giudizi di valutatori esperti. Tuttavia, ogni attività di valutazione è stata condotta separatamente, senza che chi la conduceva avesse a disposizione i punteggi assegnati nelle altre attività. Ai valutatori descritti di seguito – con l’eccezione dei valutatori esperti, responsabili anche della preparazione del campione – i testi sono poi stati sottoposti senza indicazioni sulla provenienza o sull’origine delle riformulazioni.

6.1. Valutatori esperti

Una prima valutazione del lavoro è stata compiuta dai due autori. Mirko Tavosanis è un ricercatore attivo da oltre 25 anni nel settore della chiarezza comunicativa; ha pubblicato in proposito un manuale scritto in collaborazione [16] e contributi divulgativi e scientifici dedicati alla valutazione dei testi generati. Mariachiara Pascucci è dottoranda presso la Scuola di Dottorato in Italianistica dell’Università di Pisa con una ricerca sul miglioramento della chiarezza nella comunicazione amministrativa.

In una prima fase, i due valutatori hanno lavorato in modo indipendente. I punteggi da loro assegnati sono stati poi confrontati per produrre una valutazione condivisa, che è stata usata come punto di riferimento.

6.2. Valutatori formati appositamente

Il gruppo è stato composto da studenti frequentanti del corso di Linguistica italiana II del corso di laurea magistrale in Informatica umanistica dell’Università di Pisa. Il corso di laurea richiede alle matricole il possesso di almeno 12 CFU in discipline linguistiche all’ingresso; diversi studenti hanno poi competenze più avanzate negli studi linguistici.

Tutti i valutatori hanno quindi operato mentre seguivano un corso annuale sulla valutazione dei testi generati. La sezione conclusiva del corso è stata dedicata alla valutazione del miglioramento della chiarezza, con l’inclusione di basi teoriche, la descrizione dei tratti linguistici tipicamente coinvolti e una formazione specifica sulla valutazione. Al termine del corso si è svolta un’attività di armonizzazione delle valutazioni in presenza (90 minuti), in cui le valutazioni assegnate a testi simili a quelli poi presi in esame sono state discusse e revisionate in modo da arrivare a una valutazione quanto più possibile condivisa.

L’attività finale di valutazione è stata svolta in presenza, in aula, con testi presentati su carta e una durata di 90 minuti. I valutatori sono stati divisi in due gruppi, denominati A (7 valutatori) e B (6 valutatori); ogni gruppo doveva valutare 8 testi riformulati, 4 dei quali prodotti da ChatGPT e 4 da intervento umano, accompagnati dagli originali; i testi erano alternati nei due gruppi, in modo che nel complesso venissero valutati tutti gli 8 testi prodotti da ChatGPT e tutti gli 8 prodotti da intervento umano. Non tutti i valutatori hanno completato l’attività, in particolare per gli ultimi testi di ogni gruppo.

6.3. Crowdsourcing

I testi sono stati valutati anche mediante crowdsourcing, utilizzando la piattaforma Prolific.

L'uso di metodi di crowdsourcing per la ricerca linguistica è ben documentato, come descritto in [17]. In particolare, sistemi di crowdsourcing sono stati applicati anche al campo della complessità linguistica e del miglioramento della chiarezza in lavori come [3], [11] e [18].

Per questo lavoro, la selezione dei partecipanti è stata realizzata avviando due studi distinti per ottenere due gruppi differenziati di valutatori. I criteri di selezione includevano la padronanza della lingua italiana e il livello di istruzione; sono stati infatti reclutati solo partecipanti in possesso di un diploma di laurea.

Per replicare le condizioni della valutazione in aula, è stato reclutato lo stesso numero di partecipanti, suddivisi in Gruppo A (7 valutatori) e Gruppo B (6 valutatori).

Il tempo a disposizione per completare l'attività era identico a quello della valutazione in aula, ovvero 90 minuti, ma il tempo impiegato in media dai partecipanti per lo svolgimento del compito è stato di 35 minuti. I gruppi di testi distribuiti ai partecipanti su Prolific corrispondevano, per ordine e tipologie di rielaborazione, a quelli utilizzati nella valutazione in aula. Prolific ha reindirizzato i partecipanti selezionati a un modulo Google. Nella scheda iniziale del modulo sono state fornite le indicazioni per l'assegnazione dei punteggi, identiche a quelle fornite per la valutazione in aula. Ogni scheda successiva del modulo conteneva il testo originale e la versione revisionata, con l'istruzione di assegnare un punteggio da 1 a 5 per ciascuno dei parametri specificati.

6.4. Valutazione con ChatGPT

L'attività di valutazione è stata condotta anche con ChatGPT (versione 3.5), proponendo come prompt al sistema le stesse istruzioni fornite ai valutatori umani. ChatGPT è stato impiegato in modalità zero-shot: per lo svolgimento del compito non sono dunque stati forniti al modello esempi di valutazioni già realizzate. Le versioni originali e quelle rielaborate di ciascun testo sono state presentate a ChatGPT separatamente in diverse finestre di dialogo, senza specificare l'origine della revisione, analogamente a quanto fatto con i valutatori umani. Pur non avendo ricevuto indicazioni specifiche a tal proposito, ChatGPT ha fornito, per ogni parametro, una motivazione dettagliata del punteggio assegnato, facendo ampio riferimento ai criteri di valutazione forniti.

7. Risultati della valutazione

Va notato che in tutti e quattro i modi la valutazione ha classificato le rielaborazioni come di alto livello. I voti assegnati ai singoli aspetti da valutare non scendono in effetti quasi mai sotto il 3 e rimangono quasi sempre nella fascia del 4 e del 5. Le differenze tra i singoli valutatori umani e ChatGPT sono quindi piuttosto contenute. La sintesi dei risultati completi è presentata nell'Appendice C.

Una discussione dei risultati in rapporto alle prestazioni del sistema viene presentata in [3] e [4]. Qui verranno invece prese in considerazione solo le differenze nei risultati tra i quattro modi di valutazione. Occorre quindi innanzitutto confrontare le medie complessive della valutazione (Tabella 1).

Tabella 1

Medie complessive e indicazione dello scostamento assoluto rispetto al valore fornito dagli esperti.

	Gruppo A	scostamento	Gruppo B	scostamento
Esperti	**4,40**		**4,66**	
Valutatori formati	4,51	0,11	4,58	0,08
Crowdsourcing	4,23	0,17	3,90	0,76
GPT	4,92	0,52	4,86	0,20

Tra i vari modi di valutazione ci sono dunque differenze rilevanti nei risultati. Usando come riferimento i giudizi dei valutatori esperti, il maggior avvicinamento si ha con i giudizi dei valutatori formati. GPT fornisce punteggi sistematicamente più alti (in pratica, tutti 5 con pochi 4), mentre il crowdsourcing fornisce valutazioni sistematicamente più basse. Calcolando lo scostamento complessivo, inteso come somma dei valori assoluti delle differenze, il risultato migliore si ha con i valutatori formati, con 0,19, seguiti a buona distanza da ChatGPT con 0,76 e dal crowdsourcing con 0,93.

Le medie complessive nascondono però una differenza tra gli aspetti. Come è stato notato dai valutatori esperti, è possibile assegnare i punteggi per gli aspetti 1, 2 e 5 in modo relativamente oggettivo, appoggiandosi a valutazioni quantitative, mentre per gli aspetti 3 e 4 è frequente l'incertezza di assegnazione tra il punteggio 4 e il punteggio 5. Sembra quindi utile valutare separatamente gli aspetti 1, 2 e 5 (Tabella 2).

Tabella 2

Medie degli aspetti 1, 2 e 5 e indicazione dello scostamento assoluto rispetto al valore fornito dagli esperti.

	Gruppo A	scostamento	Gruppo B	scostamento
Esperti	**4,42**		**4,76**	
Valutatori formati	4,58	0,16	4,54	0,18
Crowdsourcing	4,32	0,10	4,02	0,74
GPT	4,92	0,50	4,92	0,16

Anche in questo caso, calcolando lo scostamento complessivo, il risultato migliore si ha comunque con i valutatori formati, con 0,34, seguiti da ChatGPT con 0,66 e dal crowdsourcing con 0,84. La classifica quindi non cambia, anche se è notevole che su questa selezione di aspetti lo scostamento minore rispetto agli esperti si ottenga con il crowdsourcing nel gruppo A e con ChatGPT nel gruppo B.

7.1. Accordo tra valutatori

Per quanto riguarda la robustezza della valutazione sia nel caso dei valutatori formati appositamente sia nel caso del crowdsourcing, l'accordo tra i valutatori individuali non ha raggiunto i livelli considerati sufficienti secondo il calcolo dell'alpha di Krippendorff ([19]).

L'accordo complessivo tra i valutatori formati appositamente per il gruppo A è stato in effetti di 0,288; per il gruppo B, di 0,270. Il livello massimo di accordo è stato raggiunto dal gruppo A nella valutazione dell'aspetto di "conservazione delle informazioni", che ha raggiunto il valore di 0,502. L'accordo complessivo tra i valutatori reclutati per crowdsourcing è stato invece di 0,181 per il gruppo A e di 0,141 per il gruppo B. Anche in questo caso, il livello massimo di accordo è stato raggiunto dal gruppo A nella valutazione dell'aspetto di "conservazione delle informazioni", che però ha raggiunto solo il valore di 0,241.

Secondo lo schema di interpretazione dell'alpha di Krippendorff, i valori inferiori a 0,670 sono

indicative of poor agreement among raters. Data with a Krippendorff's Alpha below this threshold are often deemed unreliable for drawing triangulated conclusions. It suggests that the raters are not applying the coding scheme consistently or that the scheme itself may be flawed [19].

La conclusione è senz'altro condivisibile: per questo tipo di attività, semplicemente, i criteri devono ancora essere messi a punto in modo soddisfacente. Tuttavia, sembra evidente la distanza tra l'accordo che si può raggiungere con valutatori formati appositamente invece che con il semplice crowdsourcing.

7.2. Esame di un caso specifico

I motivi per le differenze tra le diverse valutazioni sono naturalmente molto difficili da ricostruire. Tuttavia, in almeno alcuni casi è possibile notare che i valutatori hanno fornito valutazioni difficili da giustificare oggettivamente, in rapporto probabile con la complessità del compito.

Per esempio, nel caso del testo con codice CASS-4, inserito nel gruppo A, il crowdsourcing ha fornito una valutazione di 4,71 per l'aspetto di "conservazione delle informazioni". Su sette valutatori, infatti, 4 hanno fornito il punteggio 5, che corrisponde al giudizio "la riformulazione è sostanzialmente corretta e completa"; 3 invece hanno fornito un punteggio di livello 4, che corrisponde al giudizio "la riformulazione altera l'originale, ma solo in modo marginale (per esempio, se viene descritta una procedura, le discrepanze possono portare distorsioni marginali nella procedura; rientrano in questa categoria anche le leggere ma ripetute distorsioni del significato di partenza, anche quando sono prive di effetti pratici)". I valutatori esperti hanno invece assegnato al testo il giudizio 3: "la riformulazione altera l'originale in modo significativo (per esempio, se viene descritta una procedura, anche se la maggior parte dei contenuti è corretta, seguire le indicazioni potrebbe provocare qualche errore nello svolgimento della procedura)".

Quest'ultima valutazione è stata dovuta, in aggiunta a cambiamenti minori, al modo in cui sono state riformulate due frasi.

Frase 1 – originale: [Il Consiglio Circoscrizionale] risolve, per motivi espressi in narrativa, di fissare i seguenti criteri al fine di consentire all'ufficio Tecnico circ.le di essere autorizzato ad installare direttamente ovvero di autorizzare i cittadini ad installare barriere metalliche senza dover sottoporre le richieste al parere della Commissione circ.le Traffico

Frase 1 - riformulazione: Il Consiglio Circoscrizionale, in virtù delle ragioni esposte, decide di stabilire i seguenti criteri per consentire all'Ufficio Tecnico circoscrizionale di autorizzare direttamente l'installazione dei parapedonali o di concedere ai cittadini l'autorizzazione per farlo senza dover ricorrere al parere della Commissione circoscrizionale del Traffico.
(Generated by AI tool ChatGPT-3.5)

Nel testo riformulato non si chiarisce che l'Ufficio Tecnico può installare direttamente i parapedonali, ma si parla solo di "autorizzare" l'installazione.

Frase 2 – originale: i parapedonali dovranno essere installati lungo i marciapiedi in corrispondenza delle Ambasciate che, in genere per motivi di sicurezza, ne fanno richiesta.

Frase 2 – riformulazione: Sarà richiesto l'installazione di parapedonali lungo i marciapiedi in corrispondenza delle Ambasciate, su richiesta di queste ultime per ragioni di sicurezza.
(Generated by AI tool ChatGPT-3.5)

Nel testo riformulato, al di là dell'errore grammaticale (e semantico) in "Sarà richiesto l'installazione", viene inserita una limitazione inesistente nell'originale: le installazioni possono implicitamente essere accettate solo nel caso che dietro alla richiesta vi siano "ragioni di sicurezza". Quella che nell'originale era solo una spiegazione diventa quindi una condizione necessaria.

In questo caso, i valutatori esperti confermano l'attribuzione del giudizio 3, che però non è stato espresso da nessun valutatore del crowdsourcing (nella valutazione da parte di esperti, il testo è stato valutato solo da 2 valutatori, che hanno comunque assegnato il giudizio 5).

8. Conclusioni

I risultati dei diversi modi di valutazione potrebbero a prima vista essere interpretati come una svalutazione del crowdsourcing, rispetto al quale la semplice richiesta a ChatGPT è in grado di fornire risultati di qualità più alta. Tuttavia, è chiaro che le caratteristiche dell'attività svolta rendono consigliabile non trarre conclusioni troppo generalizzate.

Innanzitutto, invita alla cautela il fatto che la valutazione dipenda con ogni evenienza dalla scala usata. In un contesto in cui si sa che il voto può essere solo 4 o 5, in fin dei conti, la semplice assegnazione casuale del punteggio darebbe 4,5 sia al gruppo A sia al gruppo B, scostandosi dal giudizio degli esperti con 0,26 per la valutazione complessiva e 0,34 per gli aspetti 1, 2 e 5, valori molto vicini a quelli forniti dai valutatori formati.

In queste circostanze, sembra innanzitutto utile creare griglie di valutazione più specifiche e mirate. Le alte prestazioni dei sistemi attuali, del resto, rendono senz'altro meno utili che in passato scale 1-5 in cui il punteggio 1 deve essere assegnato a un "testo completamente incomprensibile" e il punteggio 5 a un "testo perfettamente comprensibile".

Vanno inoltre tenuti presenti alcuni limiti dell'analisi. Uno tra questi è il coinvolgimento degli autori nella riscrittura di alcuni testi: anche se le caratteristiche della valutazione rendono a nostro giudizio molto limitato il rischio di alterazioni, si prevede di modificare il protocollo per future attività dello stesso genere, delegando tutte le riscritture a terze parti. Per la valutazione dei testi generati da ChatGPT può essere inoltre utile far valutare i testi a un sistema diverso – e, in generale, ampliare e ripetere le valutazioni è naturalmente indispensabile per validarne i risultati.

Di sicuro, però, i risultati invitano a prestare attenzione ai limiti di pratiche oggi diffuse come il crowdsourcing, che sul compito in esame hanno mostrato un notevole scostamento rispetto alla valutazione di esperti. Inoltre, se la valutazione rapida ed economica fornita da sistemi come ChatGPT dovesse essere regolarmente confermata come più vicina alla valutazione di esperti rispetto al crowdsourcing, le motivazioni per il crowdsourcing stesso scomparirebbero.

Ringraziamenti

Ringraziamo per la collaborazione Claudia Gigliotti. Diversi aspetti dell'analisi dei dati sono stati discussi con Angela Ferrari. La responsabilità delle affermazioni rimane naturalmente agli autori.

Note

[1] G. Fiorentino, V. Ganfi. "Parametri per semplificare l'italiano istituzionale: revisione della letteratura." Italiano LinguaDue 16.1, pages 220-237, 2024, doi:10.54103/2037-3597/23835

[2] M. Tavosanis, "Valutare la qualità dei testi generati in lingua italiana." AI-Linguistica 1.1 (2024), pages 1-24.

[3] W. S. Lasecki, R. Luz, J. P. Bigham, Measuring text simplification with the crowd, in: Proceedings of the 12th Web for All Conference W4A 15, 2015. doi:10.1145/2745555.2746658.

[4] M. Tavosanis, Valutare la riformulazione automatica, in: Amministrazione attiva, Firenze, Cesati (in stampa).

[5] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of Text Generation: A Survey, 2020, arXiv:2006.14799.

[6] R. Tariq et al., Assessing ChatGPT for Text Summarization, Simplification and Extraction Tasks, 2023 IEEE 11th International Conference on Healthcare Informatics (ICHI), Houston, TX, USA, 2023, pp. 746-749, 2023, doi: 10.1109/ICHI57859.2023.00136.

[7] A. Sottana, B. Liang, K. Zou, Z. Yuan, Evaluation Metrics in the Era of GPT-4: Reliably Evaluating Large Language Models on Sequence to Sequence Tasks. 2023, arXiv:2310.13800.

[8] C. van der Lee, A. Gatt, E. van Miltenburg, S. Wubben, E. Krahmer, Best practices for the human evaluation of automatically generated text. In Proceedings of the 12th International Conference on Natural Language Generation, pages 355–368, Tokyo, Japan, Association for Computational Linguistics, 2019.

[9] D. Nozza, G. Attanasio, Is It Really That Simple? Prompting Language Models for Automatic Text Simplification in Italian. CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy, 2023.

[10] N. van Raaij, D. Kolkman, K. Podoynitsyna, Clearer Governmental Communication: Text Simplification with ChatGPT Evaluated by Quantitative and Qualitative Research. In Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 152–178, Torino, Italia. ELRA and ICCL, 2024.

[11] D. Brunato, L. De Mattei, F. Dell'Orletta, B. Iavarone, G. Venturi, Is this sentence difficult? do you agree? In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2690–2699, Brussels, Belgium Association for Computational Linguistics, 2018.

[12] M. Cortelazzo, Il linguaggio amministrativo: principi e pratiche di modernizzazione, Carocci, Roma, 2021.

[13] S. Cassese (a cura di), Codice di stile delle comunicazioni scritti ad uso delle amministrazioni pubbliche, Istituto poligrafico e zecca dello Stato, Roma, 1994.

[14] E. Piemontese, Capire e farsi capire. Teorie e tecniche della scrittura controllata. Napoli: Tecnodid, 1996.

[15] K. Krippendorff, Content Analysis: An Introduction to Its Methodology, 4th edition, SAGE Publications, Los Angeles, 2019.

[16] M. Gasperetti, M. Tavosanis, Comunicare, Apogeo, Milano, 2004.

[17] R. Munro, S. Bethard, V. Kuperman, V.T. Lai, R. Melnick, C. Potts, T. Schnoebelen e H. Tily. Crowdsourcing and language studies: The new generation of linguistic data. In Proceedings of the Workshop on Creating Speech and Language Data with Amazons Mechanical Turk, pages 122–130, 2010.

[18] O. De Clercq, V. Hoste, B. Desmet e P. Van Oosten. Using the crowd for readability prediction, Natural Language Engineering, pages 1–33, 2013.

[19] G. Marzi, M. Balzano, D. Marchiori, K-Alpha Calculator—Krippendorff's Alpha, 2024. Calculator: A User-Friendly Tool for Computing Krippendorff's Alpha Inter-Rater Reliability Coefficient. MethodsX, 12, 102545, 2024, doi: https://doi.org/10.1016/j.mex.2023.102545

A. Prompt usati

Prompt 1: Puoi semplificare la forma linguistica del seguente testo amministrativo-burocratico pur mantenendo tutti i dettagli del contenuto? Voglio che il testo prodotto sia dettagliato e lungo tanto quanto il testo da semplificare che è qui tra virgolette "[...]"

Prompt 2: Rendi più chiaro il seguente testo inserito tra virgolette, estratto da linee guida ministeriali, in modo che sia facilmente comprensibile per un pubblico diversificato, inclusi individui con conoscenze limitate dell'argomento e un livello medio di istruzione. Concentrati sull'utilizzo di un linguaggio chiaro e conciso senza compromettere l'accuratezza delle informazioni. Assicurati che siano preservati i dettagli chiave riguardanti la procedura descritta. Punta a migliorare l'accessibilità e la leggibilità mantenendo il contenuto e il significato essenziali del documento. Preserva la coesione del testo. Mantieni bilanciata la lunghezza del testo. "[...]"

B. Griglia di valutazione e istruzioni

1. La correttezza delle informazioni fornite

1: la riformulazione non ha nessun rapporto con l'originale o altera l'originale (per omissione, deformazione o aggiunta) al punto di essere incomprensibile

2: la riformulazione altera l'originale in modo grave (per esempio, se viene descritta una procedura, il testo riformulato non permette di eseguirla correttamente)

3: la riformulazione altera l'originale in modo significativo (per esempio, se viene descritta una procedura, anche se la maggior parte dei contenuti è corretta, seguire le indicazioni potrebbe provocare qualche errore nello svolgimento della procedura)

4: la riformulazione altera l'originale, ma solo in modo marginale (per esempio, se viene descritta una procedura, le discrepanze possono portare distorsioni marginali nella procedura; rientrano in questa categoria anche le leggere ma ripetute distorsioni del significato di partenza, anche quando sono prive di effetti pratici)

5: la riformulazione è sostanzialmente corretta e completa

Precisazioni importanti

L'**omissione**, totale o parziale, dei riferimenti a leggi, regolamenti e simili deve essere considerata **ininfluente** (a meno che non sia necessaria per spiegare una parte del testo: per esempio, il fatto che le modifiche sono richieste da una legge appena approvata): questa va considerata come una scelta redazionale presa a monte.

Quindi per esempio dovranno essere considerate buone, dal punto di vista della correttezza delle informazioni, riformulazioni come questa:

Originale: Approvazione, con Decreto del Ministero del Lavoro e delle Politiche Sociali n. 15 del 29 gennaio 2024, della "Nota Metodologica per l'adozione di UCS (Unità di Costo Standard)."

Riformulato: Approvazione della "Nota Metodologica per l'adozione di UCS (Unità di Costo Standard)" con un decreto ministeriale del gennaio 2024.

Anche **l'omissione di informazioni** (purché non rilevanti alla comprensione di quanto rimane) deve essere considerata ininfluente: anche l'eliminazione di informazioni deve essere considerata una scelta redazionale. L'entità dell'omissione viene valutata invece nell'aspetto 5.

Quindi per esempio dovranno essere considerate buone, dal punto di vista della correttezza delle informazioni, riformulazioni come queste:

Originale: l'introduzione dell'equivalenza alla partecipazione ai PUC, ai fini della definizione degli impegni nell'ambito dei patti per l'inclusione sociale, della partecipazione, definita d'intesa con il Comune, ad attività di volontariato presso enti del Terzo settore e a titolarità degli stessi, da svolgere nel Comune di residenza nei medesimi ambiti di intervento previsti per i PUC;

Riformulato: l'introduzione dell'equivalenza tra partecipazione ai PUC e ad attività di volontariato per i patti per l'inclusione sociale.

Un buon modo per controllare può essere: dare brevi titoli ai singoli capoversi, per sintetizzare l'argomento, e valutare la correttezza un capoverso alla volta.

2. La correttezza linguistica del testo

Nella prospettiva di un lettore italiano medio (madrelingua, con diploma di scuola superiore come titolo di studio più alto), dal punto di vista formale il testo risulta:

1: difficile da ricondurre alla norma

2: con quattro o più errori morfosintattici (indipendentemente dalla loro estensione)

3: con non più di tre errori morfosintattici e/o molti usi insoliti di collocazioni, o simili

4: con non più di due errori morfosintattici, possibili anche a esseri umani, e/o non più di due usi insoliti delle collocazioni, o simili aspetti discutibili dal punto di vista formale

5: corretta, con incertezze minime che potrebbero essere trovate anche in un testo professionale umano

Precisazioni importanti

La valutazione di questo aspetto **non deve riguardare il registro linguistico**. In altri termini, la scelta di usare un tono più o meno formale, incluso l'impiego di forestierismi, viene considerata una scelta redazionale.

Per esempio, in un testo potranno essere accettabili sia "fare" sia "eseguire", senza assegnare una preferenza all'una scelta o all'altra – a parità di correttezza.

La valutazione **non deve riguardare nemmeno la comprensibilità delle parole o delle espressioni**, che è valutata separatamente nell'aspetto 3. Per esempio, a livello di correttezza linguistica possono essere accettabili sia "download" sia "scaricamento", anche se una parola è più comprensibile dell'altra.

L'accettabilità di incertezze "minime" è collegata al fatto che anche lettori L1 colti possono avere idee diverse sull'accettabilità o meno di alcune parole e costruzioni. Di qui anche l'importanza di mettersi nella prospettiva di un lettore italiano "medio".

3. La chiarezza complessiva del testo

Per un lettore italiano medio (madrelingua, con diploma di scuola superiore come titolo di studio più alto), il testo riformulato è verosimilmente:

1: incomprensibile

2: quasi del tutto incomprensibile

3: in buona parte comprensibile, ma con uno o più elementi significativi poco comprensibili

4: in buona parte comprensibile, con piccole incertezze (per esempio, sul significato esatto di una parola)

5: perfettamente comprensibile

Precisazioni importanti

Questo aspetto deve essere valutato **senza tenere conto della completezza o della correttezza oggettiva delle informazioni**, ma solo della loro coerenza interna e della loro presentazione. Inoltre, deve essere valutato senza basarsi sulla brevità o meno del testo (di cui, in sede di valutazione complessiva, si tiene conto in base alla lunghezza in parole e in caratteri dell'originale e della riformulazione).

Anche per questo aspetto, come per l'aspetto 1, l'omissione o il mantenimento dei riferimenti a leggi, regolamenti e simili devono essere considerati ininfluenti: ai fini della valutazione di questo aspetto, si suppone che i riferimenti compaiano se sono utili ai fini della comunicazione e non compaiano se sono inutili ai fini della comunicazione. Lo stesso vale per l'omissione di informazioni, che viene valutata nell'aspetto 5.

4. Il livello di miglioramento rispetto all'originale

1: il testo è molto meno chiaro dell'originale

2: il testo è sensibilmente meno chiaro dell'originale

3: il testo è tanto chiaro quanto l'originale

4: il testo è sensibilmente più chiaro dell'originale

5: il testo è molto più chiaro dell'originale

Precisazioni importanti

Anche per questo aspetto, come per l'aspetto 1, l'omissione o il mantenimento dei riferimenti a leggi, regolamenti e simili devono essere considerati ininfluenti: ai fini della valutazione di questo aspetto, si suppone che i riferimenti compaiano se sono utili ai fini della comunicazione e non compaiano se sono inutili ai fini della comunicazione. Lo stesso vale per l'omissione di informazioni, che viene valutata nell'aspetto 5.

5. La conservazione delle informazioni

1: il testo elimina più del 75% delle informazioni dell'originale

2: il testo elimina tra il 75% e il 50% delle informazioni dell'originale

3: il testo elimina tra il 50% e il 25% delle informazioni dell'originale

4: il testo elimina una parte delle informazioni dell'originale inferiore al 25%

5: il testo mantiene tutte le informazioni dell'originale

Precisazioni importanti

La valutazione deve essere una stima quantitativa. Non deve tener quindi conto dell'importanza delle informazioni eliminate, ma solo della loro quantità. Si può tenere come riferimento la lunghezza delle espressioni che presentano le informazioni eliminate.

Un buon modo per valutare la conservazione delle informazioni può essere: sottolineare nell'originale le parole o le espressioni o le frasi che non hanno riscontro nel testo riformulato e fare una stima della percentuale complessiva.

Importante! In caso di dubbio sull'aspetto cui assegnare un errore o una deviazione, la **correttezza delle informazioni** (aspetto 1) deve essere privilegiata rispetto alla correttezza linguistica (aspetto 2) e alla chiarezza complessiva (aspetto 3). In pratica, l'errore andrà contato come errore di correttezza, senza influire sulla valutazione degli altri aspetti.

Per esempio, un'espressione come "Se il Beneficiario non è lo stesso dell'esecutore dell'azione" (al posto di "Qualora il Beneficiario non coincida con il Soggetto Attuatore") dovrebbe essere valutata come errore nella correttezza, indipendentemente dai dubbi che possono venire (a seconda dei contesti) per quanto riguarda la correttezza linguistica o la chiarezza.

C. Risultati complessivi

Testo	Aspetti	Esperti	Appositamente	Crowdsourcing	ChatGPT
PRIN-4 Revisione umana	Correttezza delle informazioni	5,00	4,83	4,50	5,00
	Correttezza linguistica	5,00	4,67	4,67	4,00
	Chiarezza complessiva del testo	4,00	5,00	4,50	4,00
	Livello di miglioramento	4,00	4,67	3,83	4,00
	Conservazione delle informazioni	5,00	4,83	4,17	5,00
PRIN-4 ChatGPT	Correttezza delle informazioni	5,00	5,00	4,57	5,00
	Correttezza linguistica	5,00	5,00	4,86	5,00
	Chiarezza complessiva del testo	4,00	4,57	4,71	5,00
	Livello di miglioramento	4,00	3,57	3,86	5,00
	Conservazione delle informazioni	5,00	5,00	4,43	5,00
FP-4 Revisione umana	Correttezza delle informazioni	5,00	4,43	3,71	5,00
	Correttezza linguistica	5,00	4,86	4,43	5,00
	Chiarezza complessiva del testo	5,00	4,71	3,86	5,00
	Livello di miglioramento	5,00	4,14	3,14	5,00
	Conservazione delle informazioni	4,00	4,43	3,29	5,00
FP-4 ChatGPT	Correttezza delle informazioni	4,00	3,83	3,33	5,00
	Correttezza linguistica	5,00	4,67	3,83	5,00
	Chiarezza complessiva del testo	5,00	4,83	3,83	5,00
	Livello di miglioramento	4,00	4,50	3,33	5,00
	Conservazione delle informazioni	4,00	3,67	3,33	5,00
PONTI-1 Revisione umana	Correttezza delle informazioni	4,00	4,86	4,57	5,00
	Correttezza linguistica	5,00	4,71	4,57	5,00
	Chiarezza complessiva del testo	4,00	4,57	4,43	5,00
	Livello di miglioramento	5,00	4,57	4,00	5,00
	Conservazione delle informazioni	3,00	3,43	3,71	5,00
PONTI-1 ChatGPT	Correttezza delle informazioni	4,00	4,50	4,50	5,00
	Correttezza linguistica	5,00	4,50	4,67	5,00
	Chiarezza complessiva del testo	5,00	4,83	4,33	5,00
	Livello di miglioramento	4,00	4,33	3,00	5,00
	Conservazione delle informazioni	4,00	4,33	4,33	5,00
CASS-1 Revisione umana	Correttezza delle informazioni	5,00	2,83	3,33	5,00
	Correttezza linguistica	5,00	4,83	4,00	5,00
	Chiarezza complessiva del testo	5,00	5,00	4,17	5,00
	Livello di miglioramento	5,00	4,33	3,83	5,00
	Conservazione delle informazioni	5,00	4,67	3,50	5,00
CASS-1 ChatGPT	Correttezza delle informazioni	5,00	3,86	4,71	5,00
	Correttezza linguistica	5,00	4,00	4,57	5,00
	Chiarezza complessiva del testo	5,00	4,00	5,00	5,00
	Livello di miglioramento	4,00	3,86	5,00	5,00
	Conservazione delle informazioni	5,00	4,14	4,71	5,00

Testo	Aspetti	Esperti	Appositamente	Crowdsourcing	ChatGPT
MOB-1 Revisione umana	Correttezza delle informazioni	4,00	4,57	4,29	5,00
	Correttezza linguistica	5,00	5,00	4,57	5,00
	Chiarezza complessiva del testo	4,00	4,57	4,29	5,00
	Livello di miglioramento	4,00	4,43	3,57	5,00
	Conservazione delle informazioni	4,00	4,00	4,14	5,00
MOB-1 ChatGPT	Correttezza delle informazioni	5,00	4,67	4,00	5,00
	Correttezza linguistica	5,00	5,00	4,33	5,00
	Chiarezza complessiva del testo	5,00	4,83	4,17	5,00
	Livello di miglioramento	4,00	4,50	3,33	5,00
	Conservazione delle informazioni	5,00	4,83	3,67	5,00
ENERG-2 Revisione umana	Correttezza delle informazioni	5,00	3,67	4,33	5,00
	Correttezza linguistica	5,00	4,17	4,50	5,00
	Chiarezza complessiva del testo	5,00	3,83	4,50	5,00
	Livello di miglioramento	4,00	3,83	4,17	5,00
	Conservazione delle informazioni	5,00	4,00	4,17	5.00
ENERG-2 ChatGPT	Correttezza delle informazioni	3,00	3,57	4,29	5,00
	Correttezza linguistica	4,00	4,00	4,43	5,00
	Chiarezza complessiva del testo	5,00	4,00	4,14	5,00
	Livello di miglioramento	4,00	3,29	4,00	5,00
	Conservazione delle informazioni	4,00	4,00	4,29	5,00
PRIN-5 Revisione umana	Correttezza delle informazioni	5,00	4,57	4,14	4,00
	Correttezza linguistica	5,00	4,71	4,71	5,00
	Chiarezza complessiva del testo	4,00	4,86	4,29	5,00
	Livello di miglioramento	4,00	4,00	3,57	5,00
	Conservazione delle informazioni	3,00	3,57	3,29	5,00
PRIN-5 ChatGPT	Correttezza delle informazioni	5,00	4,00	4,17	5,00
	Correttezza linguistica	5,00	3,83	4,50	5,00
	Chiarezza complessiva del testo	5,00	3,83	3,50	5,00
	Livello di miglioramento	4,00	3,33	2,33	5,00
	Conservazione delle informazioni	5,00	4,00	3,83	5,00
CASS-4 Revisione umana	Correttezza delle informazioni	3,00	3,00	4,00	4,00
	Correttezza linguistica	5,00	3,33	3,67	5,00
	Chiarezza complessiva del testo	4,00	3,33	3,67	5,00
	Livello di miglioramento	5,00	3,00	3,17	4,00
	Conservazione delle informazioni	5,00	3,00	3,17	5,00
CASS-4 ChatGPT	Correttezza delle informazioni	3,00	2,14	4,57	4,00
	Correttezza linguistica	5,00	2,00	4,57	5,00
	Chiarezza complessiva del testo	5,00	2,14	4,14	5,00
	Livello di miglioramento	4,00	1,71	3,43	4,00
	Conservazione delle informazioni	5,00	2,00	4,29	5,00