



UNIVERSITÀ
DEGLI STUDI
DEL MOLISE



DIPARTIMENTO
DI SCIENZE UMANISTICHE
SOCIALI E DELLA FORMAZIONE



SEMPLE-IT: un modello di intelligenza artificiale per la semplificazione dell'italiano

Vittorio Ganfi – Marco Russodivito

Università degli studi del Molise

Indice



1. Semplificazione e accessibilità
2. Il corpus Italst
3. Introduzione alle ATS
4. Descrizione delle fasi di lavoro
5. I modelli per la semplificazione automatica
6. Il software SEMPL-IT
7. La valutazione

ON ANALYTIC VERBS, COMPLEXITY, SYNTHETIC VERBS, AND SIMPLICITY
FOR ACCESSIBILITY

VERBACKSS

- (tra gli altri, Fioritto 1997, Piemontese 1991, 2023, Cortelazzo, 2014, Cortellazzo e Pellegrino 2003)



DIPARTIMENTO
DI SCIENZE UMANISTICHE
SOCIALI E DELLA FORMAZIONE



UNIVERSITÀ
DEGLI STUDI
DEL MOLISE

Semplificazione e accessibilità



- Ricognizione degli studi sulla semplificazione.
- Due modalità espressive possono inficiare la leggibilità dei testi amministrativi:
 - A. azione offuscatrice:** scegliere equivalente semantico meno accessibile (ad esempio *al fine di/per*)
 - B. azione incrementale:** scegliere struttura più articolata sul piano strutturale (ad esempio ipotassi/ paratassi)

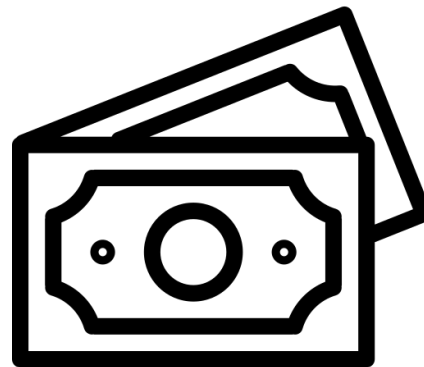
Possono essere individuati dei caratteri ricorrenti nella semplificazione:

- 13 parametri di complessità morfosintattici
 - 8 parametri di complessità lessicale
- (Fiorentino e Ganfi, in stampa)

Semplificazione e accessibilità




- La semplificazione manuale di testi è un'operazione che richiede molto tempo
 - Semplificazione di 8 documenti (619 paragrafi, 33.297 token) ha richiesto mediamente 20 ore di lavoro
- Problemi: Richiede persone esperte della lingua italiana, costi piuttosto alti e tempi lunghi.



ON ANALYTIC VERBS, COMPLEXITY, SYNTHETIC VERBS, AND SIMPLICITY
FOR ACCESSIBILITY

VERBACKSS

- 
- Complesso
- semplice

Semplificazione automatica

- Necessità di uno strumento automatico per la semplificazione dei testi.
 - Testo complesso in input e testo semplificato in output

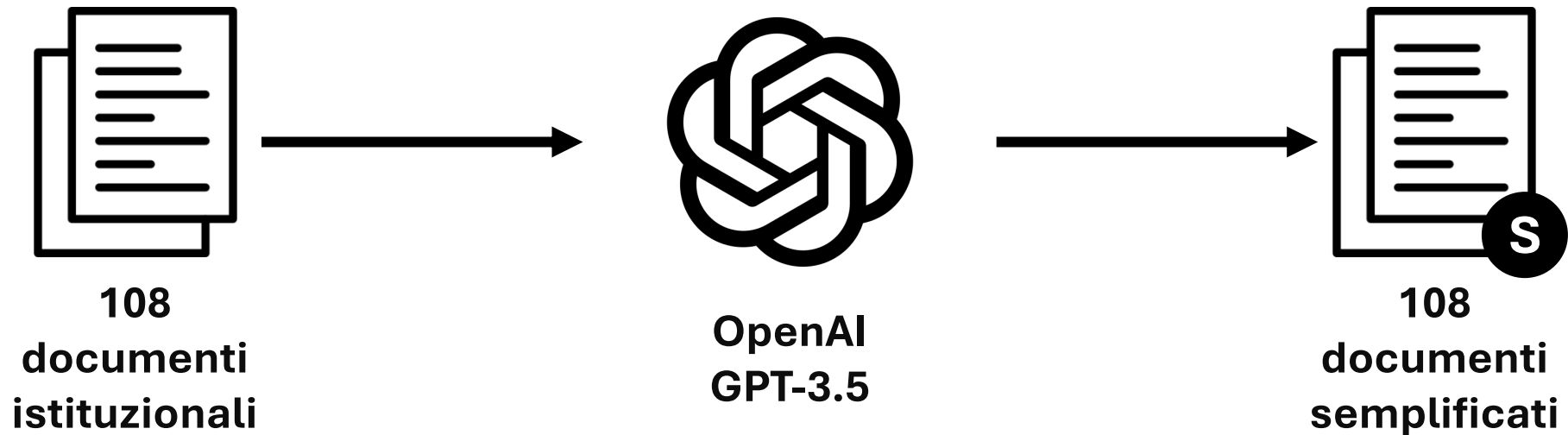


- CAVEAT: Non sostituiscono l'umano, ma forniscono uno strumento
 1. per velocizzare il processo di semplificazione
 2. per spiegare un testo complesso al fruitore che non comprende il linguaggio burocratico)

Corpus Italst



Corpus Parallelo Italst



Corpus Parallelo Italst



~ 18 mila capoversi

	Token	Type	Gulpease	VdB
Testi istituzionali	~ 840 mila	~ 24 mila	~ 40	~ 75 %
Testi semplificati	~ 720 mila	~18 mila	~ 48	~ 80 %

Similarità Semantica
~ 84 %

Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa.
Campobasso 23-25 maggio 2024



Corpus Parallelo Italst – Token



	Carta dei servizi rifiuti	Atti generali di prog.	Carte dei servizi pubblici	Razion. partecip. pub.	Accreditamento
Basilicata -	10 K	4 K	18 K	4 K	24 K
Calabria -	16 K	15 K			31 K
Campania -			57 K	18 K	
Lazio -	16 K			65 K	
Lombardia -		4 K	18 K	45 K	
Molise -	25 K	21 K	65 K	84 K	31 K
Toscana -				28 K	72 K
Veneto -		25 K	77 K	28 K	

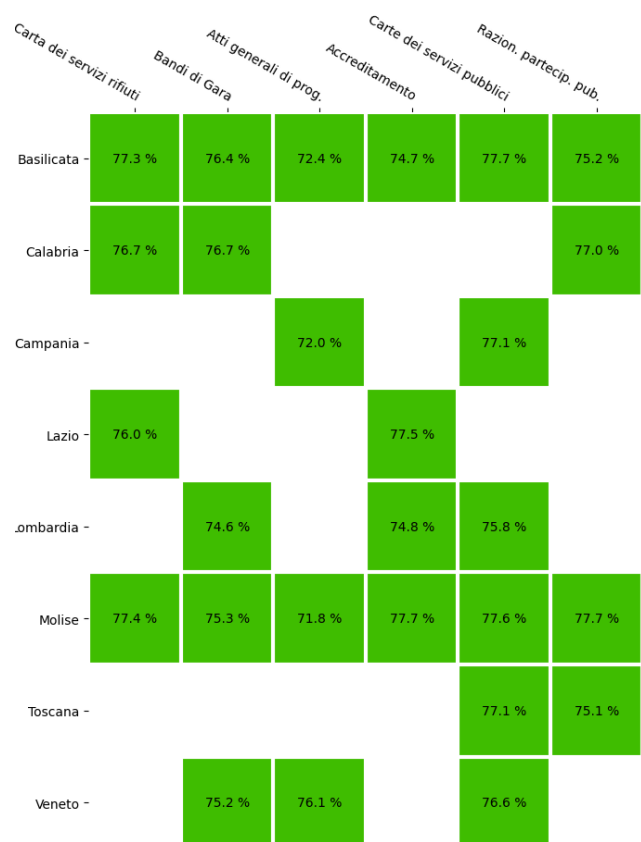
Testi istituzionali

	Carta dei servizi rifiuti	Atti generali di prog.	Carte dei servizi pubblici	Razion. partecip. pub.	Accreditamento
Basilicata -	7 K	3 K	15 K	4 K	20 K
Calabria -	13 K	13 K			24 K
Campania -			51 K	16 K	
Lazio -	14 K			56 K	
Lombardia -		3 K	14 K	39 K	
Molise -	21 K	18 K	56 K	74 K	27 K
Toscana -				23 K	58 K
Veneto -		21 K	63 K	23 K	

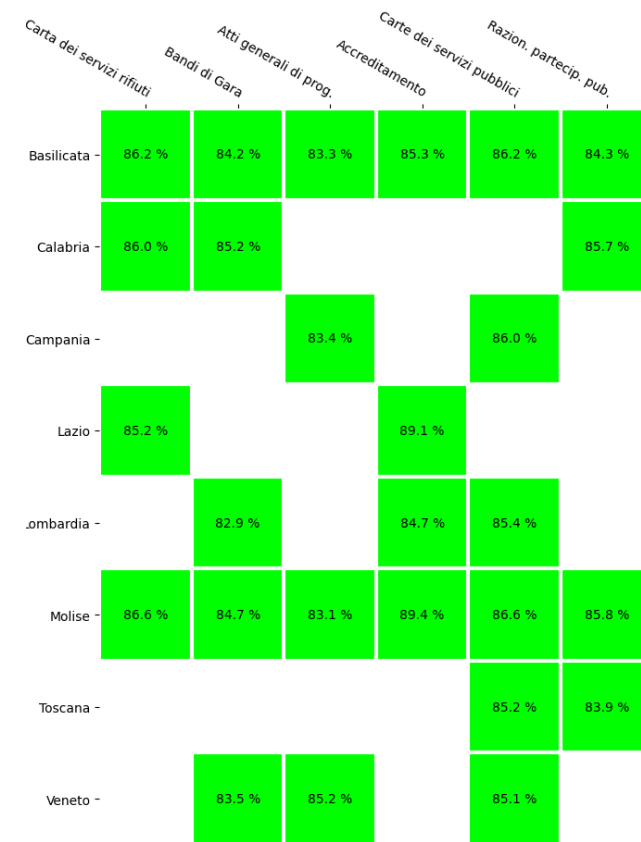
Testi semplificati

Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa.
Campobasso 23-25 maggio 2024

Corpus Parallelo Italst – VdB



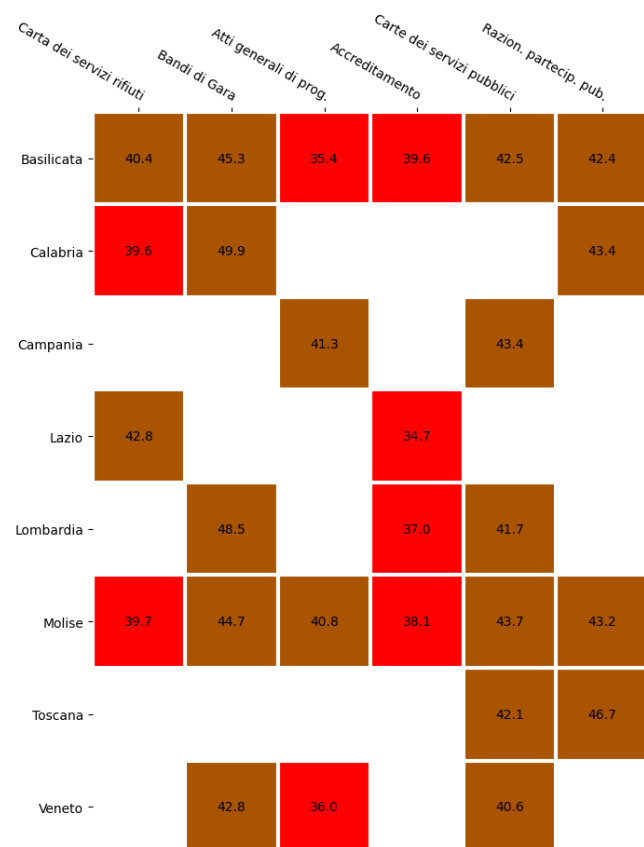
Testi istituzionali



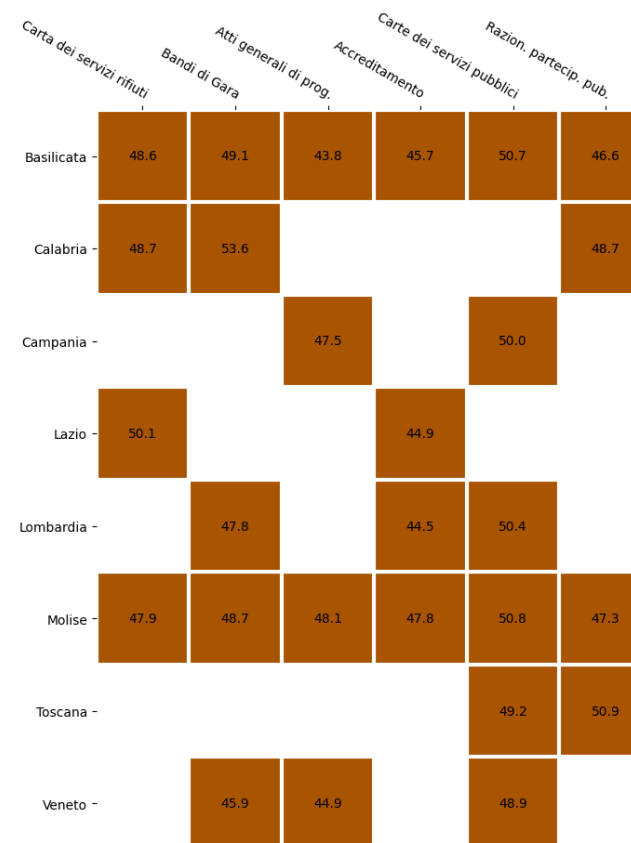
Testi semplificati

Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa.
Campobasso 23-25 maggio 2024

Corpus Parallelo Italst – Indice Gulpease



Testi istituzionali



Testi semplificati

Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa.
Campobasso 23-25 maggio 2024

Obiettivi del nostro studio



- Costruzione di un corpus parallelo
- Training di un modello ATS istituzionale
- Benchmark per valutare la qualità del ATS
- Distribuzione del modello ATS (software web)

Fasi nello sviluppo di SEMPL-IT



Corpus

- Individuazione del genere testuale (lingua amministrativa)
- Creazione di una risorsa testuale rappresentativa
- Creazione di un corpus parallelo



Semplificazione

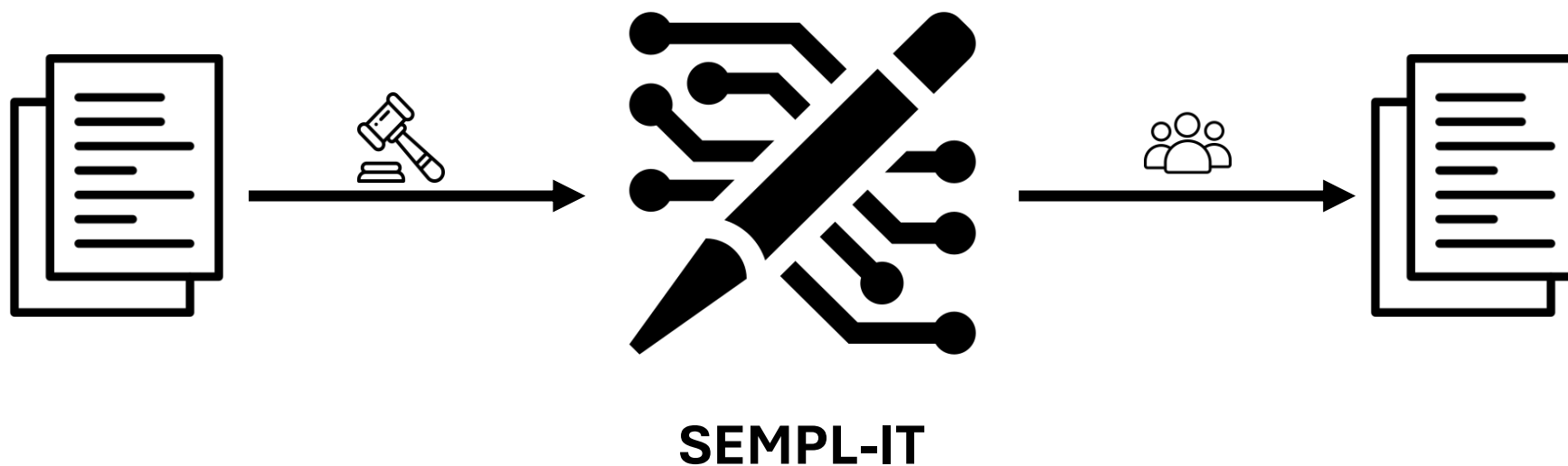
- Semplificazione del corpus mediante AI
- Semplificazione manuale di una porzione del corpus
- Confronto tra AI e umano



Sviluppo della ATS

- Addestramento di vari modelli di AI (mT5, umT5, GPT-2 ita)
- Impiegando i corpora i paralleli

SEMPLE-IT: un ATS per documenti






Cos'è l'AI Generativa?

ON ANALYTIC VERBS, COMPLEXITY, SYNTHETIC VERBS, AND SIMPLICITY
FOR ACCESSIBILITY

VERBACKSS



98 %

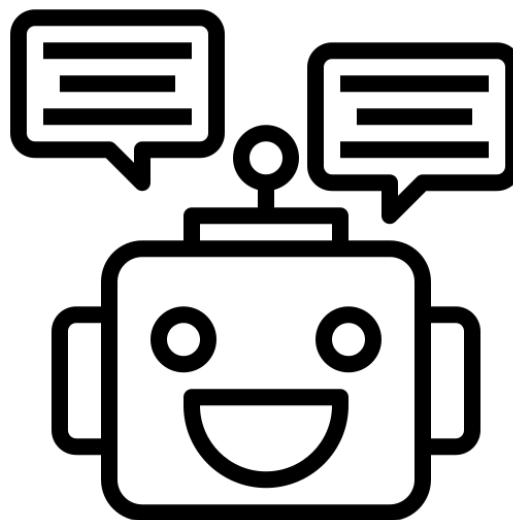
...

1%

AI Generativa – Chatbot



Ciao come stai?

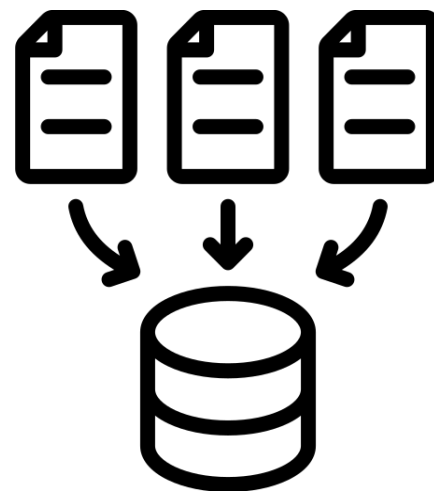


Bene grazie, tu?



Come possiamo creare un AI Generativa?

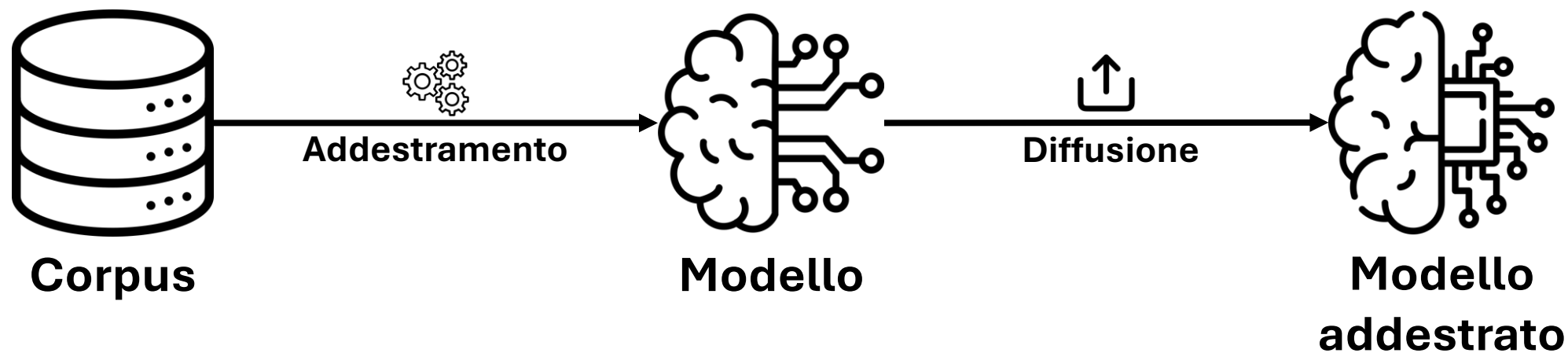
AI Generativa – Addestramento



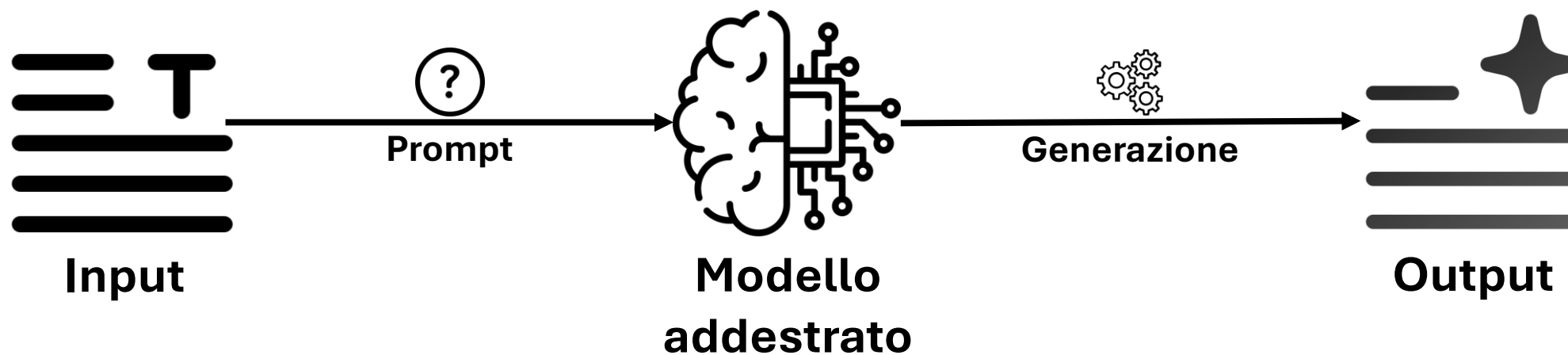
Corpus

(libri, articoli scientifici,
giornali, web ...)

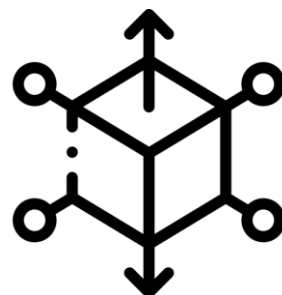
AI Generativa – Addestramento



AI Generativa – Addestramento

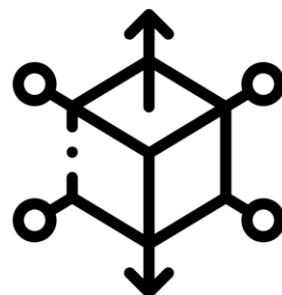


AI Generativa – Sfide



**Modelli
sempre
più grandi**

AI Generativa – Sfide

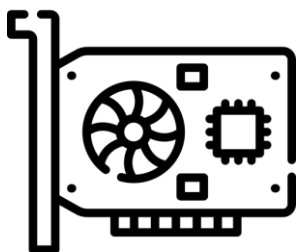


**Modelli
sempre
più grandi**

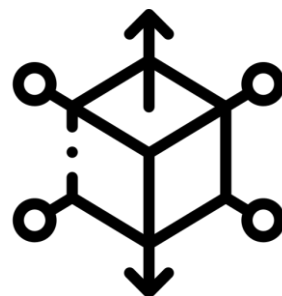


Corpus

AI Generativa – Sfide



Potenza



**Modelli
sempre
più grandi**

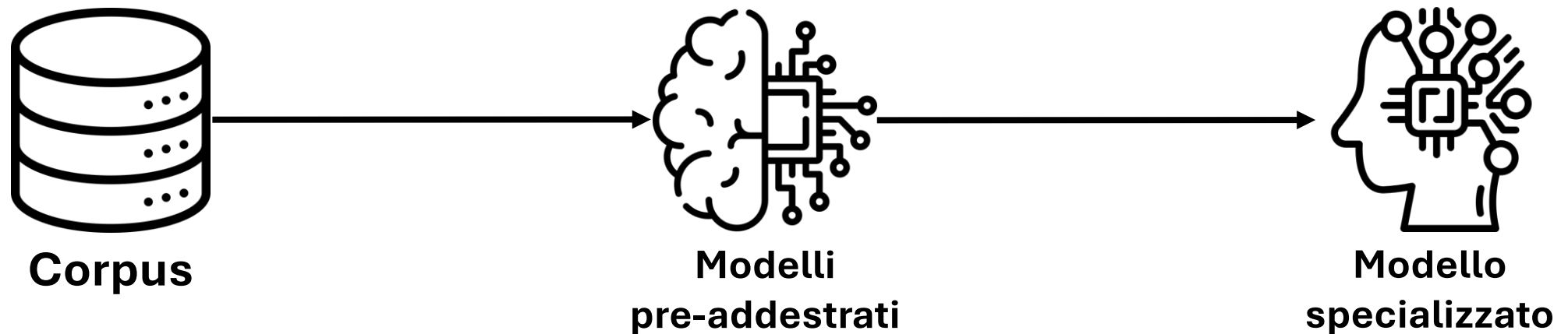


Corpus



Come abbiamo addestrato SEMPL-IT?

Specializzazione AI Generativa



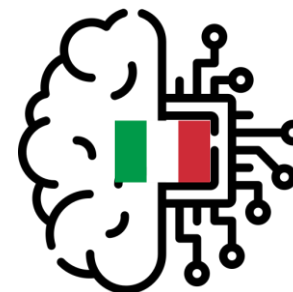
FINETUNIG

Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa.
Campobasso 23-25 maggio 2024

Addestramento SEMPL-IT

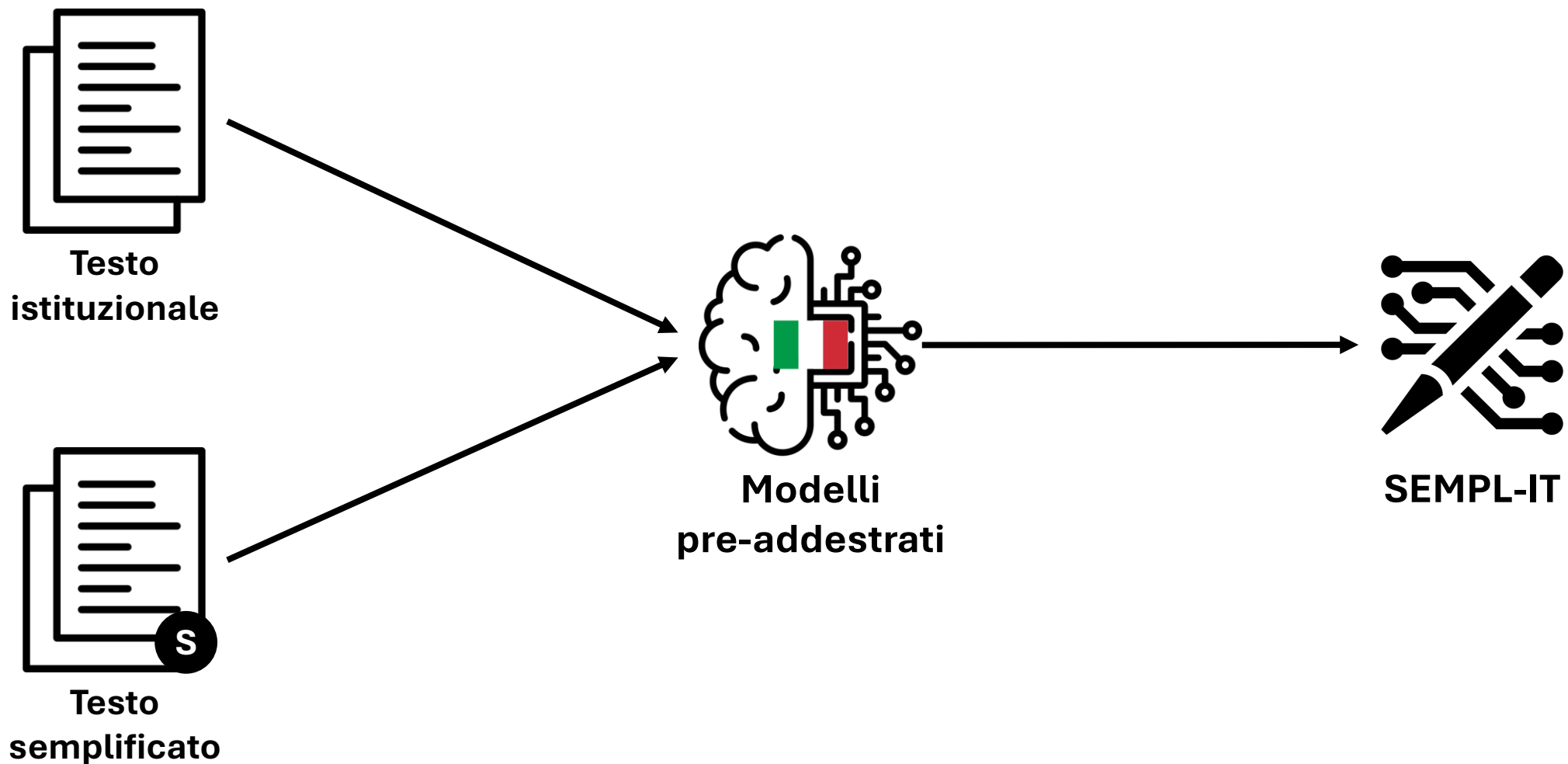


**Corpus
parallelo
Italst**



**Modelli
pre-addestrati**

Addestramento SEMPL-IT – Finetuning



Addestramento SEMPL-IT – Modello

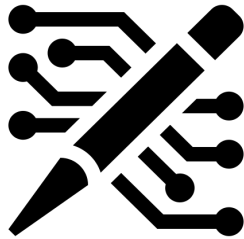


Hugging Face

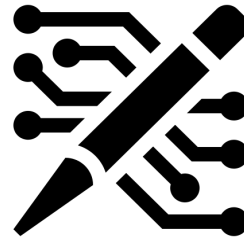


mT5
umT5
GPT-2 ITA

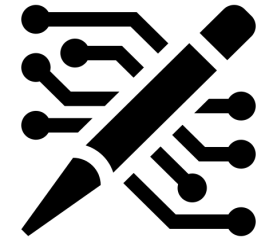
Modelli SEMPL-IT



**SEMPLE-IT
mT5**



**SEMPLE-IT
umT5**

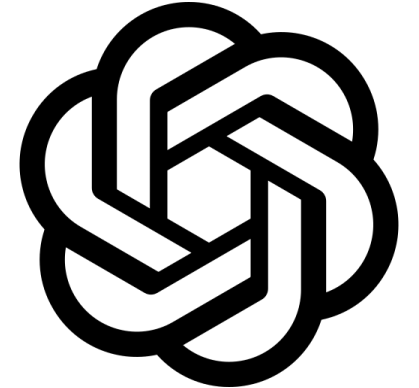
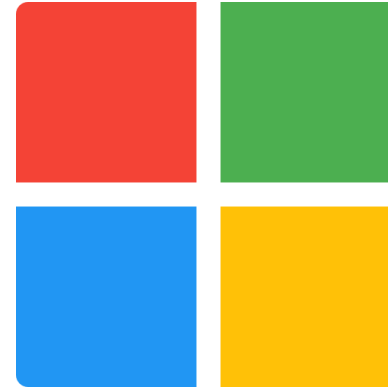
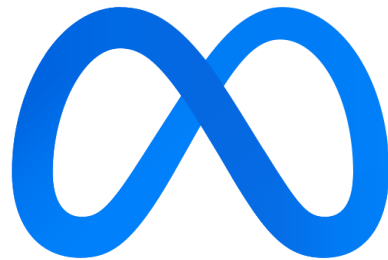


**SEMPLE-IT
GPT2 - ITA**

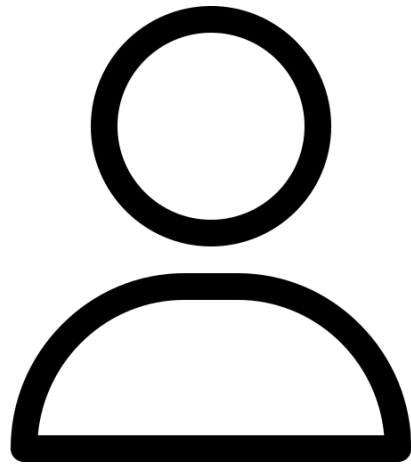


Come possiamo valutare SEMPL-IT?

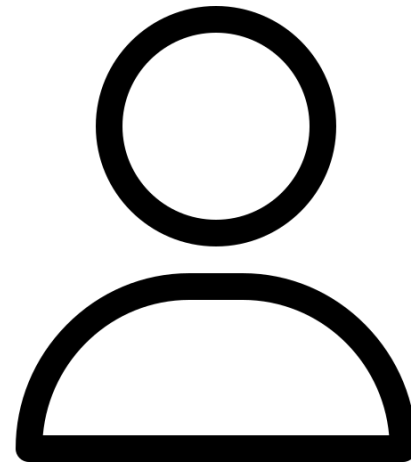
Confronto con AI popolari



Confronto con Umani



Umano 1

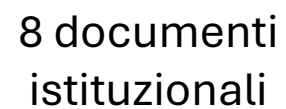


Umano 2

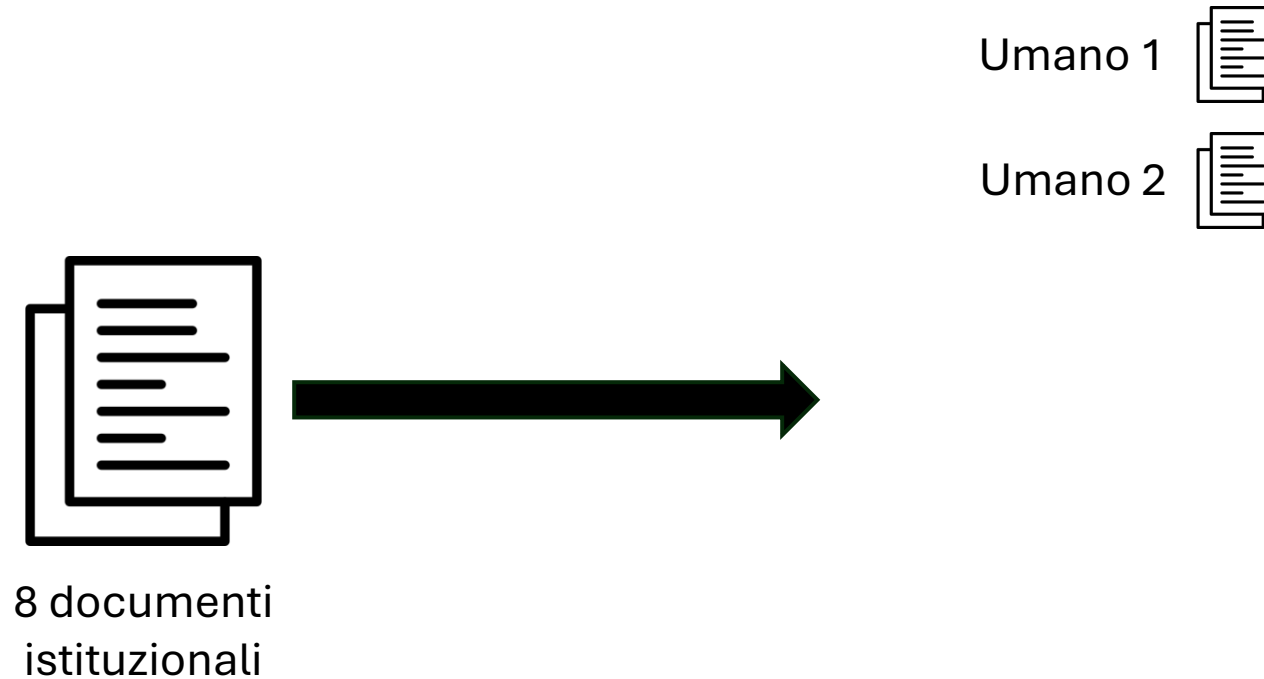
Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa.
Campobasso 23-25 maggio 2024

ON ANALYTIC VERSUS COMPLEXITY, SYNTHETIC VERSUS SIMPLIFICATION, FOR ACCESSIBILITY

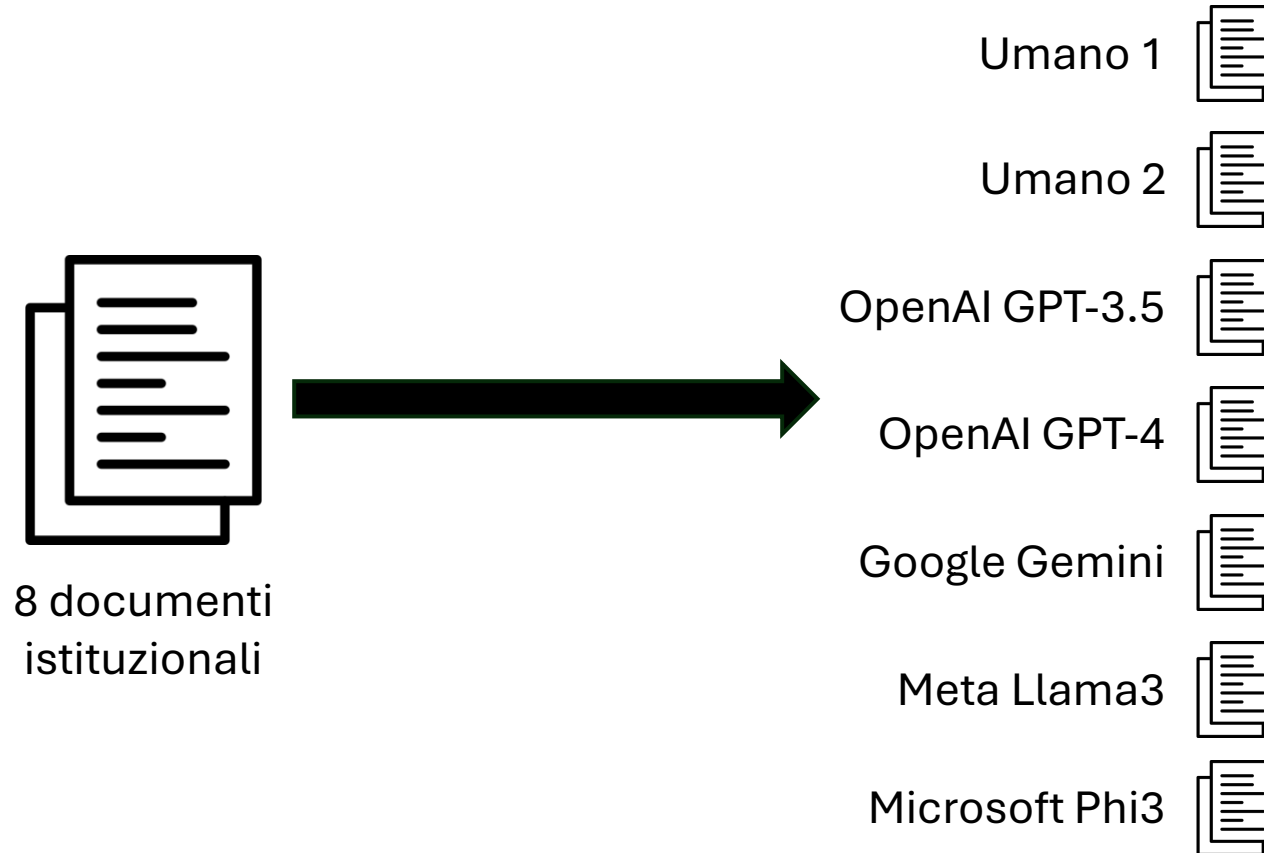
VERBACSSS



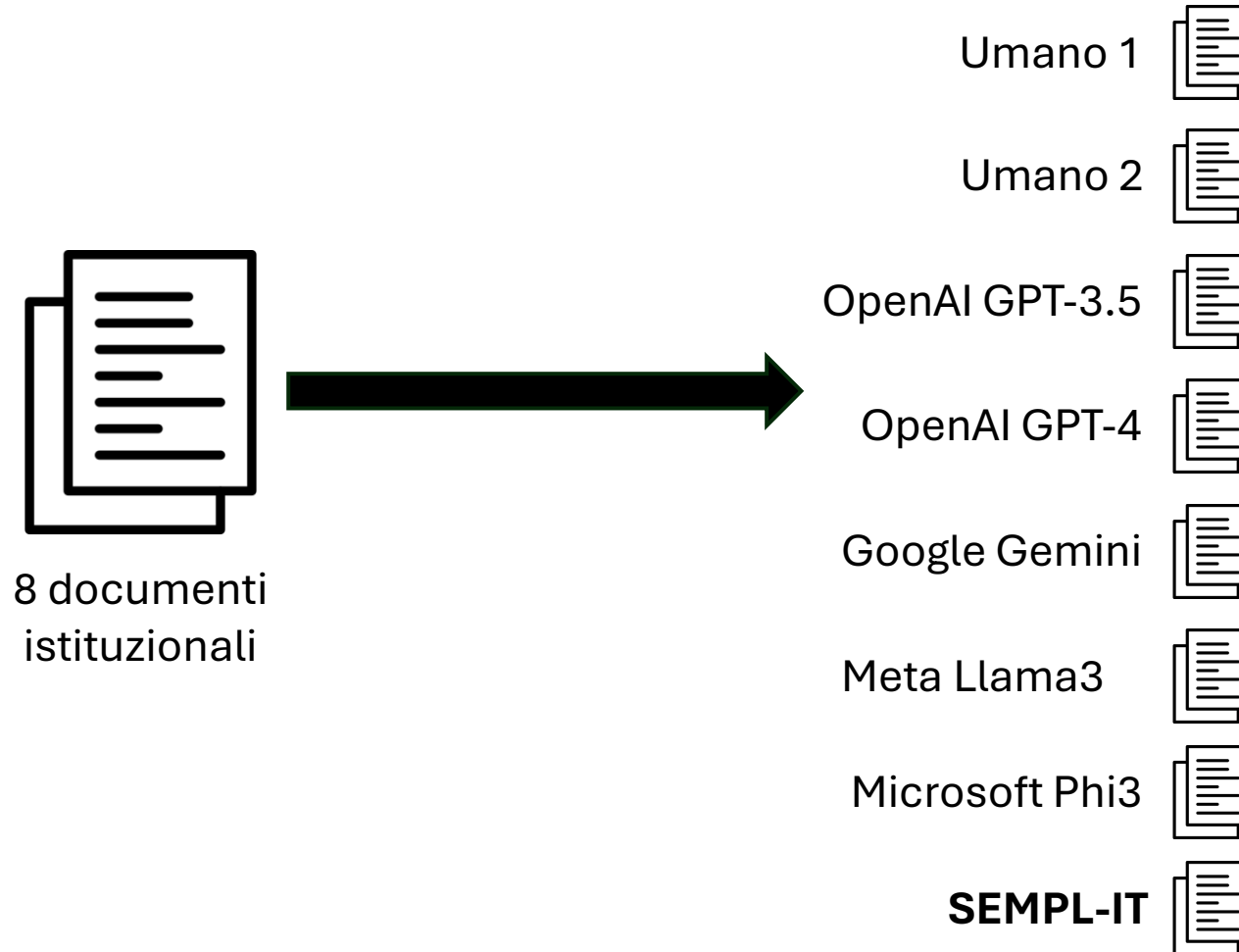
Schema di validazione



Schema di validazione

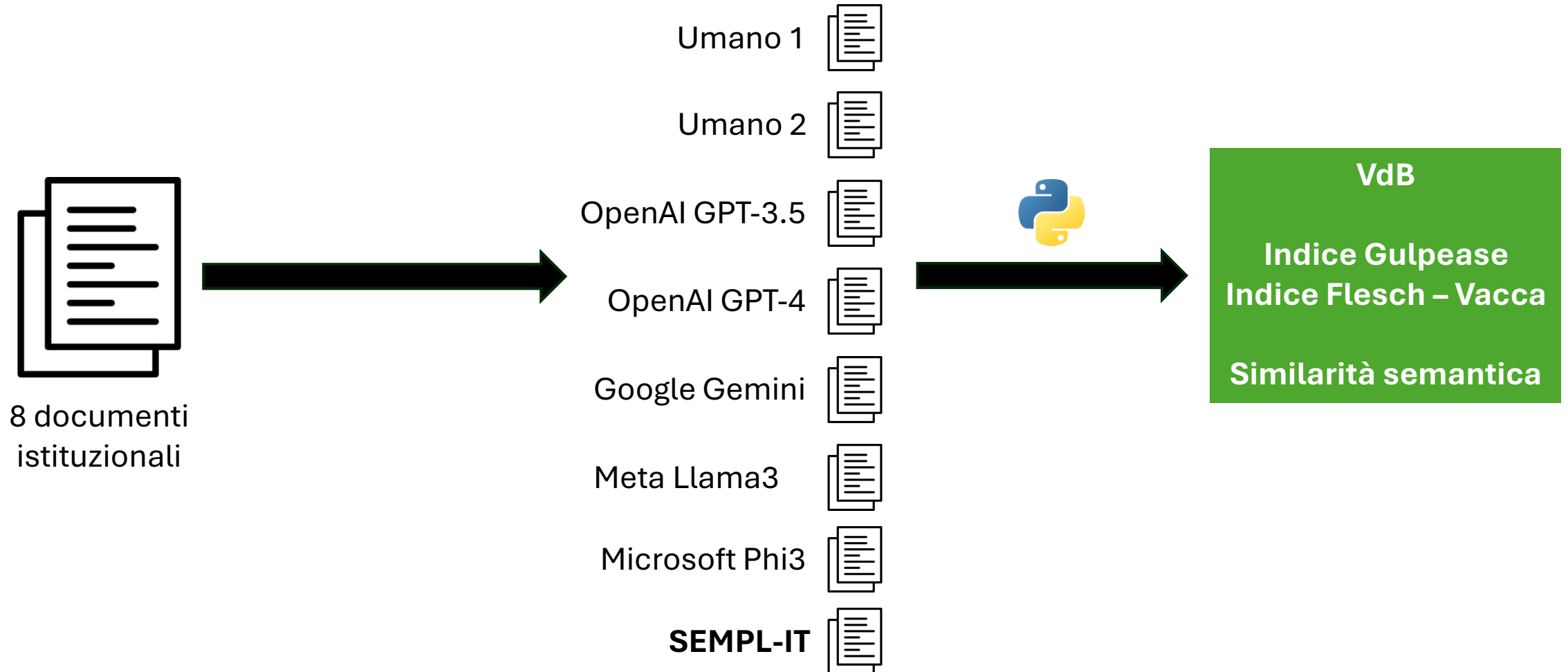


Schema di validazione



Amministrazione attiva: semplicità e chiarezza per la comunicazione amministrativa.
Campobasso 23-25 maggio 2024

Schema di validazione



Risultati



	VdB	Indice Gulpease	Indice Flesch – Vacca	Similarità semantica
Testo originale	~ 72 %	~44	~ 20	-
Umano 1	~ 79%	~49	~ 34	~ 83 %
Umano 2	~ 76 %	~ 50	~ 33	~ 87 %
OpenAI GPT-3.5	~ 77 %	~ 48	~ 30	~ 81 %
OpenAI GPT-4	~ 80 %	~ 51	~ 36	~ 80 %
Google Gemini	~ 78 %	~ 50	~ 33	~ 79 %
Meta Llama3	~ 80 %	~ 50	~ 34	~ 79 %
Microsoft Phi3	~ 80 %	~ 50	~ 33	~ 79 %
SEMPLE-IT mT5	~ 80 %	~ 50	~ 35	~ 76 %
SEMPLE-IT umT5	~ 80 %	~ 48	~ 31	~ 75 %
SEMPLE-IT GPT-2 ITA	~ 81 %	~ 50	~ 37	~ 72 %

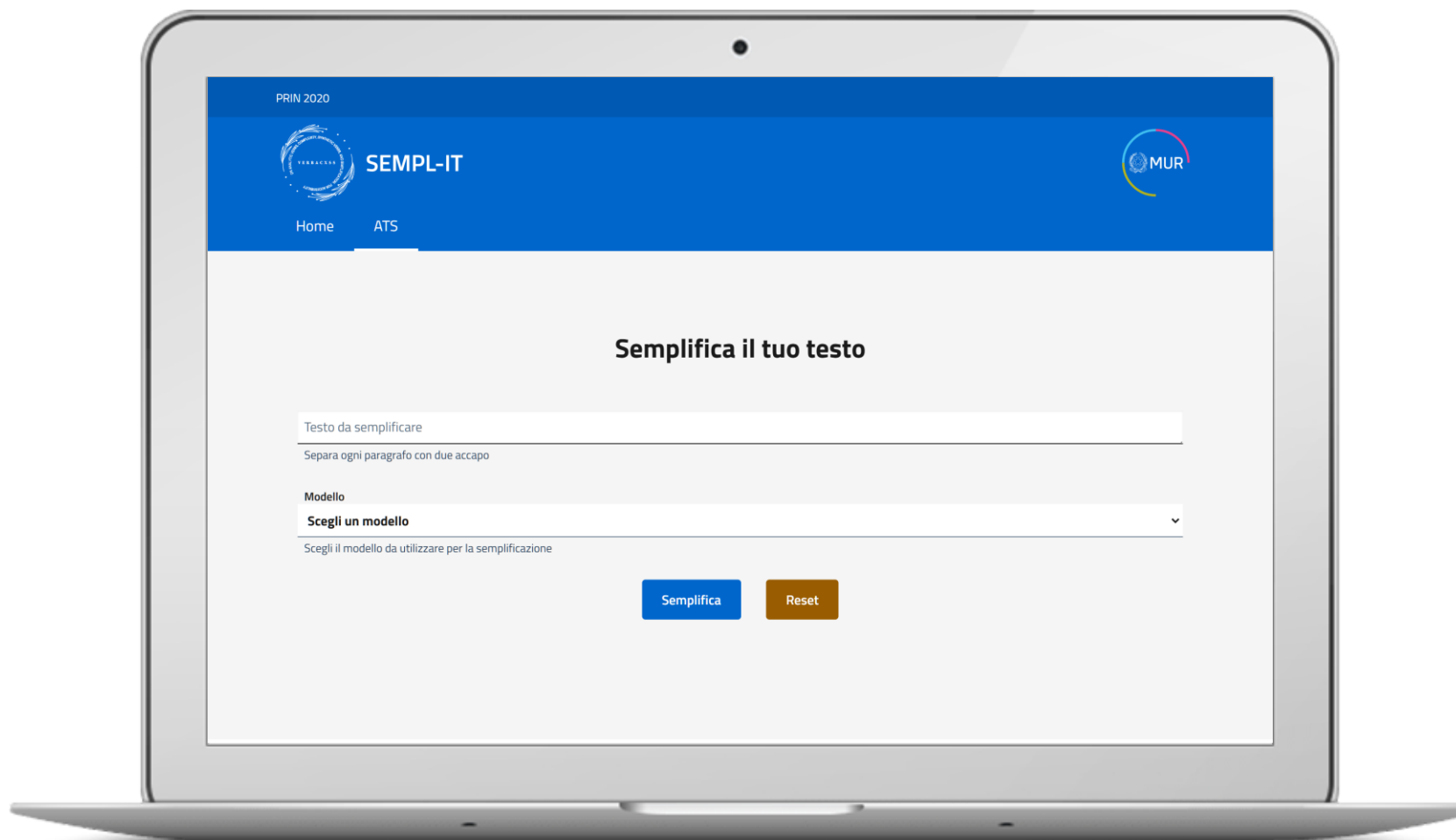


Come possiamo utilizzare SEMPL-IT?

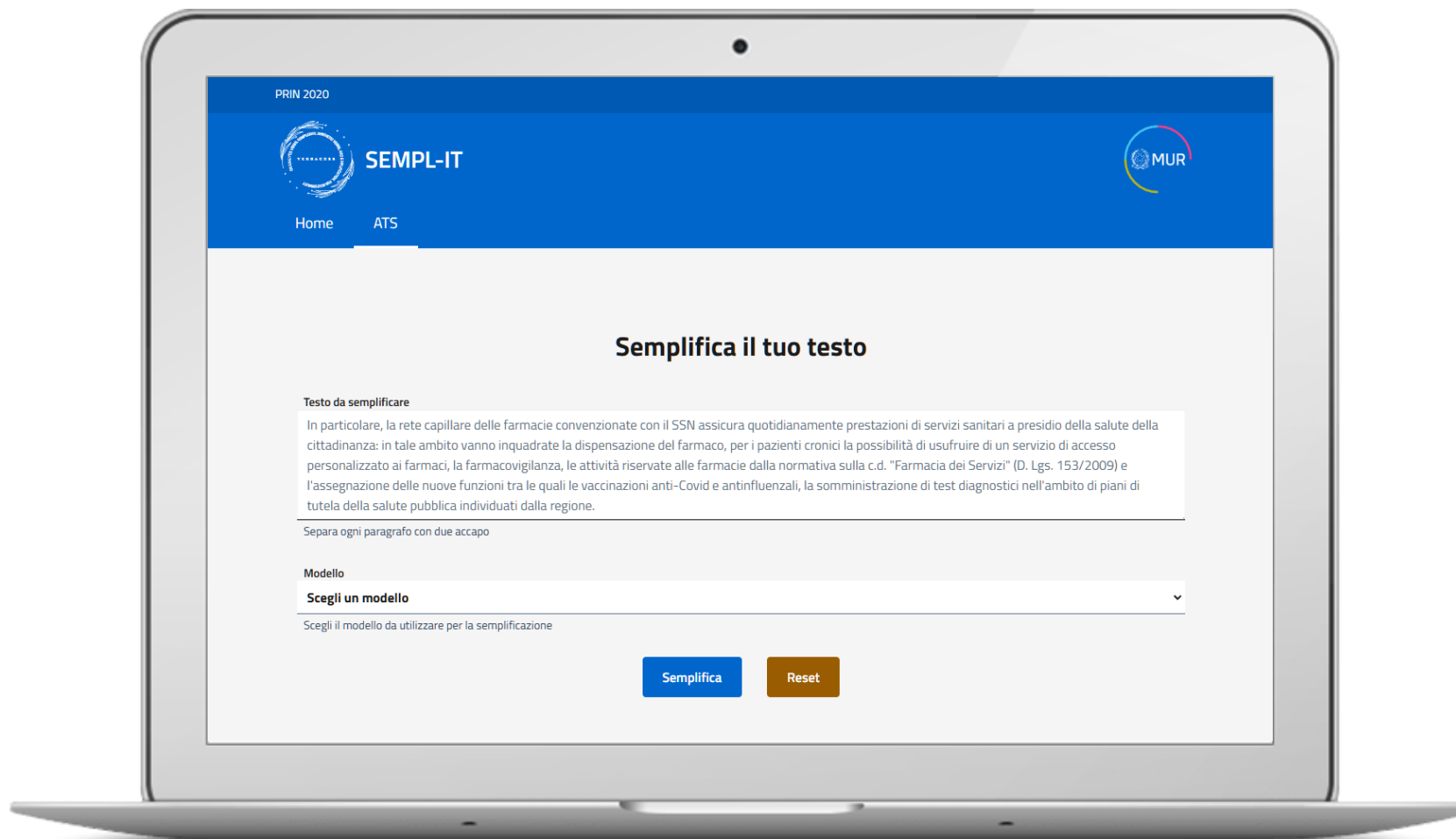
SEMPLE-IT Web



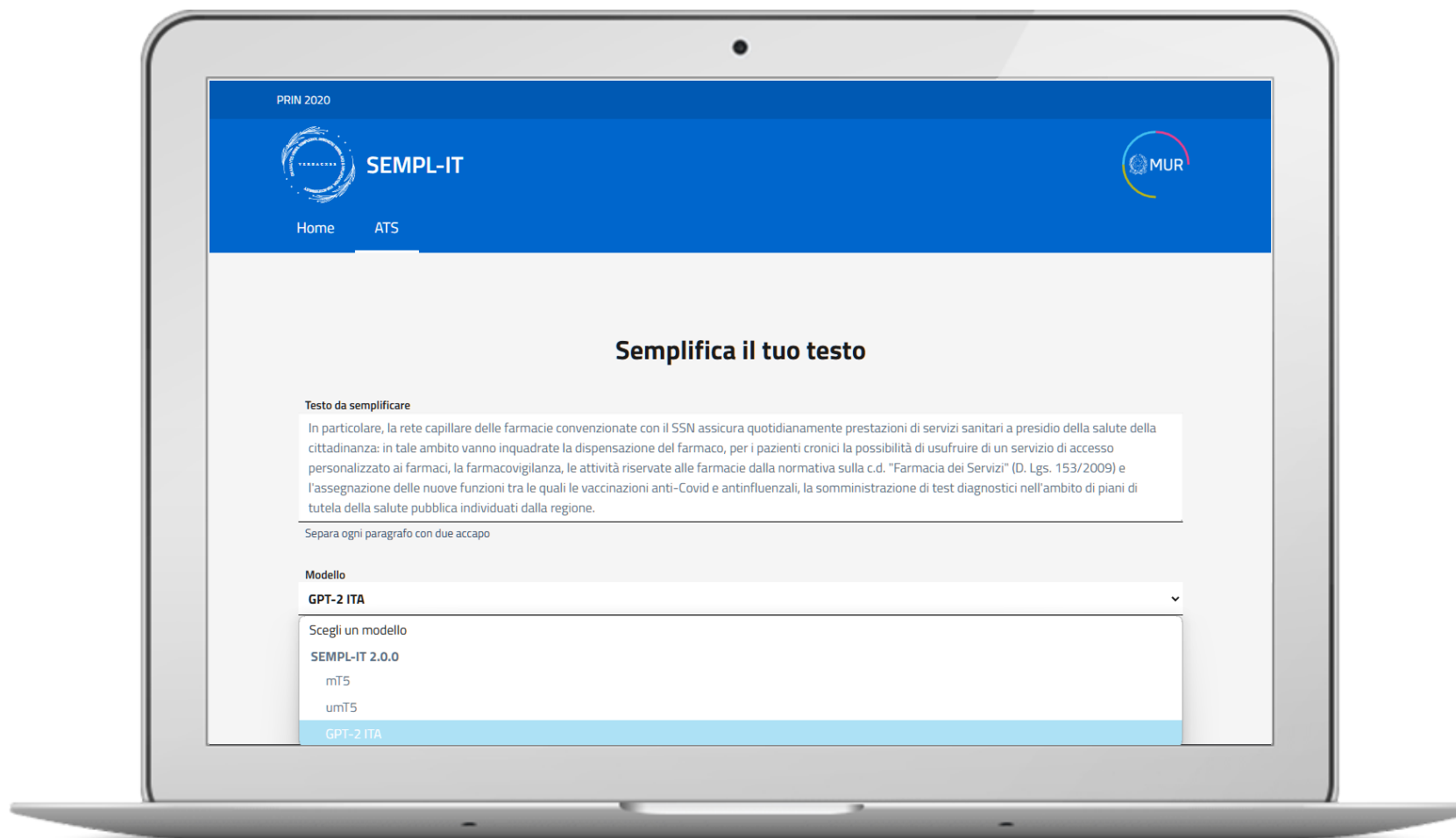
SEMPLE-IT Web



SEMP-IT Web



SEMPLE-IT Web

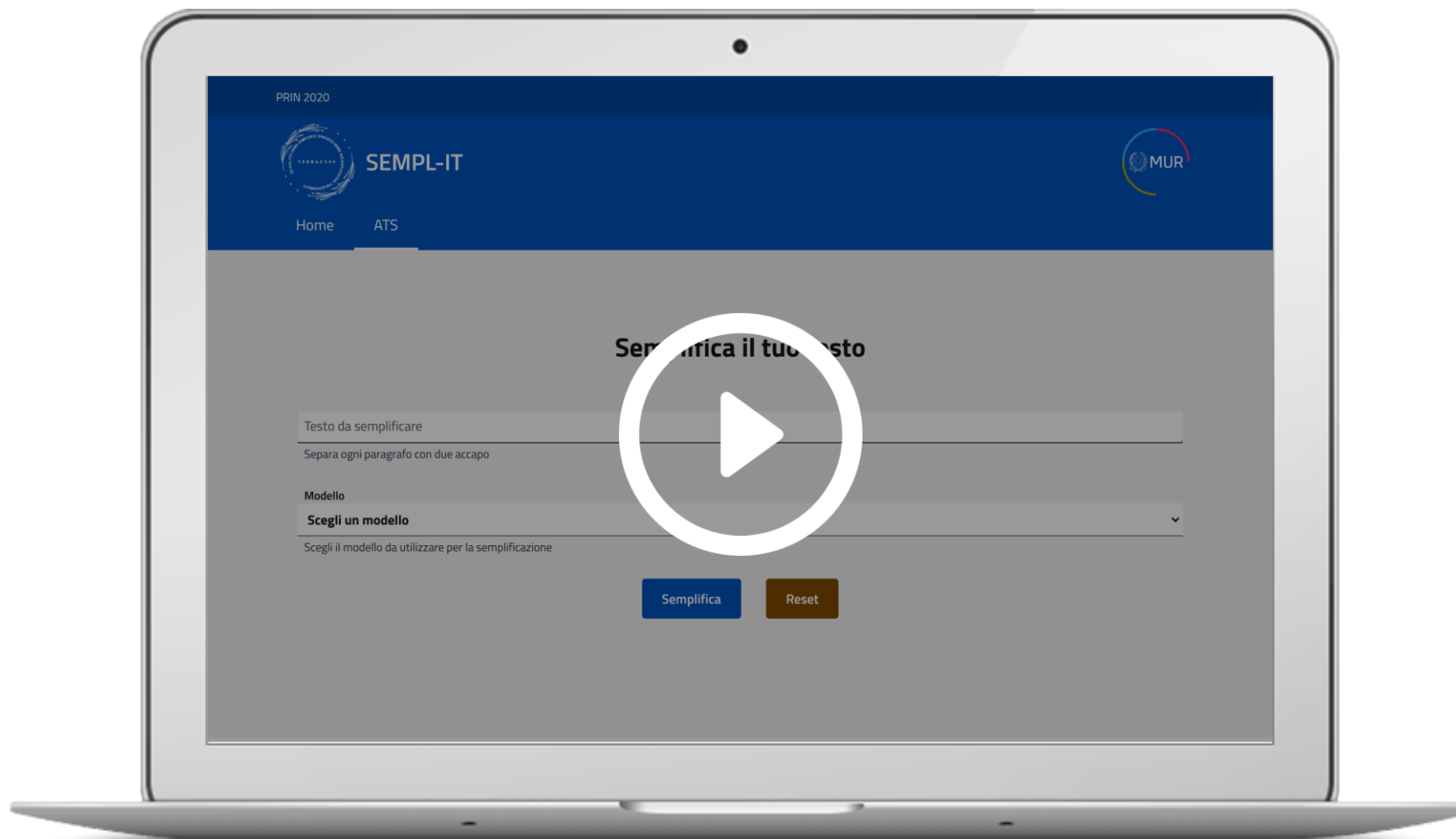


ON ANALYTIC VERBS, COMPLEXITY, SYNTHETIC VERBS, AND SIMPLICITY
FOR ACCESSIBILITY

VERBACKSS



SEMPLE-IT Web – Demo



Conclusioni

