

# La sommarizzazione delle sentenze tributarie di merito nel progetto PRODIGIT

Giuseppe Pisano<sup>1</sup> Alessia Fidelangeli<sup>1</sup> Federico Galli<sup>1</sup> Andrea Loreggia<sup>2</sup>  
Riccardo Rovatti<sup>1</sup> Piera Santin<sup>3</sup> Giovanni Sartor<sup>1</sup>

<sup>1</sup> Alma Mater Università di Bologna (Italia)

<sup>2</sup> Università degli Studi di Brescia (Italia)

<sup>3</sup> Istituto Universitario Europeo (Italia)

## Acknowledgements

Questa ricerca è stata finanziata dal Programma PON Governance e Capacità Istituzionale 2014-2020 nell'ambito del progetto "PRO.DI.GI.T.". Inoltre, la ricerca è supportata dal Consiglio europeo della ricerca (ERC) nell'ambito del programma di ricerca e innovazione Horizon 2020 dell'Unione europea (Grant Agreement n. 833647).

**Abstract:** Nel presente contributo si illustra la sommarizzazione sperimentale delle sentenze tributarie condotta nell'ambito del progetto PRODIGIT. Dopo una breve introduzione sui concetti di sommarizzazione e massimazione, si presenta la metodologia adottata nel progetto, che include la sperimentazione e la valutazione comparativa di tecniche per la sommarizzazione estrattiva e astrattiva. Si descrivono quindi i risultati ottenuti con tali tecniche, soffermandosi sulla sommarizzazione astrattiva mediante modelli linguistici di ampie dimensioni (cd. "large language models"), la tecnica che ha dato i risultati migliori. Si presentano le diverse combinazioni di comandi sperimentate (*prompt*), e i relativi risultati. Dalla valutazione, da parte di esperti tributaristi, è emerso che i sommari più corretti e completi sono forniti dalla "sommarizzazione combinata", che richiede la generazione autonoma dei diversi contenuti del sommario e la loro combinazione in un unico testo. Infine, si presenta l'uso dei sommari per funzioni di ricerca semantica.

**Parole chiave:** Diritto tributario, intelligenza artificiale, sommarizzazione astrattiva, sommarizzazione estrattiva, modelli linguistici di ampie dimensioni.

## 1 Introduzione

Da tempo, in Italia, esiste un problema di accessibilità e conoscibilità della giurisprudenza tributaria di merito. Le sentenze delle Corti di Giustizia Tributaria (già Commissioni Tributarie) non sono digitalizzate, se non per una piccola parte, e, in ogni caso, non sono sempre accessibili a professionisti e cittadini. La mancata digitalizzazione delle sentenze e una loro adeguata indicizzazione ne rendono difficile il reperimento,

✉ [g.pisano@unibo.it](mailto:g.pisano@unibo.it) (Giuseppe Pisano); [alessia.fidelangeli2@unibo.it](mailto:alessia.fidelangeli2@unibo.it) (Alessia Fidelangeli); [federico.galli7@unibo.it](mailto:federico.galli7@unibo.it) (Federico Galli); [andrea.loreggia@unibs.it](mailto:andrea.loreggia@unibs.it) (Andrea Loreggia); [riccardo.rovatti@unibo.it](mailto:riccardo.rovatti@unibo.it) (Riccardo Rovatti); [piera.santin@eui.eu](mailto:piera.santin@eui.eu) (Piera Santin); [giovanni.sartor@unibo.it](mailto:giovanni.sartor@unibo.it) (Giovanni Sartor);

🌐 <https://orcid.org/0000-0003-0230-8212> (Giuseppe Pisano); <https://orcid.org/0000-0002-8281-4176> (Alessia Fidelangeli); <https://orcid.org/0000-0003-3739-5387> (Federico Galli); <https://orcid.org/0000-0002-9846-0157> (Andrea Loreggia); <https://orcid.org/0000-0002-4731-7860> (Riccardo Rovatti); <https://orcid.org/0000-0002-0734-9657> (Piera Santin); <https://orcid.org/0000-0003-2210-0398> (Giovanni Sartor);

anche agli stessi giudici tributari. Le poche sentenze digitalizzate sono quelle rese disponibili al pubblico e acquistano una particolare rilevanza per il dibattito giuridico, a prescindere dai loro meriti sostanziali o dalla coerenza col panorama interpretativo nazionale. Il quadro attuale è reso ancora più complesso dal fatto che le diverse Corti territoriali risolvono in maniera diversa questioni interpretative tra loro analoghe (come è apparso in modo evidente, tra tutti, nel caso dell'autonoma imponibilità delle clausole penali ai fini dell'imposta di registro), così come interpretazioni diverse si susseguono in periodi di tempo limitati, pur all'interno della medesima giurisdizione territoriale. La mancanza di accessibilità della giurisprudenza limita la circolazione delle informazioni anche all'interno della magistratura tributaria, con un impatto negativo sia in termini di consapevolezza del contesto decisorio, sia di impoverimento del dibattito giurisprudenziale e accademico sui temi decisi dai giudici di merito.

In futuro si prospetta uno scenario molto diverso da quello appena presentato. È infatti in corso presso il Ministero dell'Economia e delle Finanze la predisposizione di una banca dati della giurisprudenza tributaria di merito, accessibile al pubblico. Tale banca dati accoglierà le nuove sentenze delle Corti di Giustizia Tributaria di primo e secondo grado e progressivamente si estenderà a quelle pregresse, fino a comprendere tutte le pronunce di merito.

Quando la banca dati sarà disponibile, si passerà improvvisamente dalla scarsità alla sovrabbondanza delle informazioni disponibili. Giudici e avvocati avranno accesso a tutta la giurisprudenza, ma sarà difficile, in assenza di strumenti adeguati, estrarre dalla massa delle sentenze quelle rilevanti per il caso da affrontare.

Il progetto PRODIGIT si è posto l'obiettivo di anticipare questa esigenza, usando l'Intelligenza artificiale (IA) per facilitare l'individuazione delle sentenze rilevanti e la comprensione del loro contenuto.<sup>1</sup>

L'obiettivo principale di PRODIGIT è quindi quello di offrire al giurista la possibilità di operare una sorta di "trriage" delle sentenze potenzialmente rilevanti, così da selezionare quelle che sono significative per il caso in esame e ricostruire il quadro interpretativo in cui una specifica questione giuridica deve essere inquadrata. Questa funzionalità è particolarmente utile per i giudici, che potrebbero verificare l'esistenza di filoni interpretativi precedenti consolidati o meno. In secondo luogo, potrebbe essere utile ai professionisti, che potrebbero fondare le loro richieste sulla base di decisioni già assunte in precedenza, anche da Corti diverse da quella adita.

Il procedimento di selezione della giurisprudenza proposto da PRODIGIT si basa su due pilastri: (i) la ricerca semantica delle sentenze, sulla base di quesiti formulati in testo libero, più precisa e completa possibile, così da garantire una selezione affidabile delle sentenze proposte all'esito della ricerca; (ii) un sistema di redazione di brevi sommari (sommarizzazione) del contenuto di ciascuna sentenza, così da agevolare l'utente nella scelta se procedere o meno alla lettura integrale della stessa. La presenza di funzioni per la ricerca semantica e di sommari all'interno di una banca dati della giurisprudenza di merito consentirebbe di fornire informazioni rilevanti e facilmente fruibili all'utente della banca dati, sia esso giudice, avvocato o semplice cittadino, così da garantire al contempo conoscibilità ed efficienza.

A tal fine, abbiamo innanzitutto proceduto a un approfondito esame dello stato dell'arte delle tecnologie informatiche per la ricerca e la sommarizzazione delle sentenze. Sulla base di questa indagine abbiamo sperimentato varie soluzioni prototipali. Fin dall'inizio, l'obiettivo è stato quello di sviluppare modelli operativi e scalabili (applicabili anche ad una grande quantità di sentenze), che potessero fornire la base per realizzazioni operative, immediatamente fruibili dagli operatori del processo tributario e dai cittadini.

L'ideale che ha ispirato tali modelli è stato quello di una IA non intesa a sostituire il giurista, bensì a sostenerlo nella sua attività. Si è adottata, in altri termini, una prospettiva c.d. "umano-centrica", che concepisce l'IA come strumento di ampliamento e fruizione agevole della conoscenza, riservando all'uomo le scelte e le decisioni rilevanti. Per poter creare dei prototipi rispondenti a questo obiettivo abbiamo creato un gruppo

---

1. PRODIGIT è un progetto finanziato con fondi a carico del Programma PON Governance e Capacità Istituzionale 2014-2020, Asse 7, Obiettivo specifico 7.2, Azione 7.2.1, e prevede una serie di attività per l'innovazione della Giustizia Tributaria, con il supporto della tecnologia digitale e della intelligenza artificiale. Il progetto, a cui partecipano il Dipartimento delle finanze (DF) e il Consiglio di Presidenza della giustizia tributaria (CPGT), ha ottenuto il finanziamento europeo in data 25 maggio 2022, con la stipula tra l'Agenzia per la coesione territoriale e il Dipartimento, quale Beneficiario, della Convenzione per la regolamentazione dei rapporti di attuazione, gestione e controllo relativi al PRODIGIT.

di lavoro interdisciplinare, composto di esperti di informatica, diritto tributario e informatica giuridica (para. 4.1). In questo modo, la progettazione delle soluzioni tecnologiche è stata ispirata fin dall'inizio alle esigenze della pratica del diritto. Inoltre, nella fase di validazione, sono stati coinvolti nel gruppo di lavoro diversi soggetti (*stakeholder*), rappresentanti delle diverse istanze potenzialmente interessate agli esiti del progetto (giudici tributari, ministero, università, ecc.) (para. 4.3). Abbiamo considerato le diverse tecnologie disponibili, vagliando ciascuna di esse alla luce dei risultati sperimentali (para. 4.4).

Nel presente contributo esporremo dapprima le fasi di sviluppo di uno strumento di IA per la sommarizzazione delle sentenze tributarie. Poi illustreremo brevemente la possibilità di utilizzare i sommari al fine della ricerca semantica.

La scelta di sommarizzare tutte le sentenze di merito<sup>2</sup>, a prescindere dal loro valore e dalla loro importanza, persegue un obiettivo diverso e complementare rispetto alla massimazione delle sole sentenze più rilevanti (tipica della nostra tradizione giuridica). Tale scelta è giustificata dalla possibilità — solo oggi disponibile — di applicare tecniche automatiche per la generazione di sommari, suscettibili di essere applicate a tutte le sentenze. Per meglio comprendere le ragioni di questa scelta è però necessario introdurre brevemente il concetto e la funzione di una “massima” nell’ordinamento italiano e, d’altro canto, la via che abbiamo scelto di percorrere col ricorso alla sommarizzazione.

Innanzitutto, è opportuno precisare che le massime, nella nostra cultura giuridica, sono enunciazioni sintetiche atte a cogliere i principi giuridici più importanti di una pronuncia di legittimità il cui contenuto sia di particolare interesse e che esse sono generalmente poste a fondamento della pronuncia stessa<sup>3</sup>. Nell’ordinamento nazionale, le massime prototipiche sono quelle prodotte dall’Ufficio Massimario della Suprema Corte di Cassazione<sup>4</sup> in relazione a pronunce della Corte medesima, e corrispondono alla funzione nomofilattica di quest’ultima, cioè all’obiettivo di “proteggere il diritto” assicurandone la coerente interpretazione<sup>5</sup>.

È per questa ragione che il processo di massimazione ufficiale, ad opera della stessa Corte di Cassazione, non riguarda tutte le sentenze dei giudici di legittimità, ma solo le più rilevanti sotto il profilo della nomofilachia. In particolare, i criteri per la massimazione sono cinque: (i) risoluzione da parte delle Sezioni Unite di un contrasto di giurisprudenza o di una questione di massima particolarmente importante; (ii) novità del principio; (iii) difformità rispetto ai precedenti; (iv) utilità della conferma del principio, vuoi per la sua rilevanza, vuoi per il tempo trascorso dalla più recente enunciazione o vuoi per la sua fruibilità in futuri casi analoghi; (v) rilevanza della fattispecie, per il particolare impatto sociale della questione<sup>6</sup>. È chiaro come l’obiettivo dell’intervento di massimazione sia appunto quello nomofilattico, con una duplice conseguenza: l’interesse è integralmente concentrato sul principio di diritto affermato, non sulle circostanze fattuali e, al contempo, l’Ufficio del Massimario è autorizzato a manipolare il testo per garantire la migliore comprensibilità esterna di detto principio interpretativo. La creazione di una massima richiede una profonda comprensione del rilievo della sentenza nel contesto del diritto vigente, e la capacità di tradurre tale comprensione in una precisa formulazione, che possa agevolare e indirizzare l’interpretazione e l’applicazione del diritto. A tal fine è necessario un considerevole impegno di elevatissime capacità di analisi e sintesi giuridica, come quelle dei magistrati dell’Ufficio Massimario. Ciò contribuisce a spiegare perché l’attività di massimazione non possa che limitarsi a una piccola parte delle sentenze pronunciate ogni anno nel nostro Paese.

2. La sommarizzazione di tutte le sentenze di merito è l’obiettivo a cui si dovrebbe tendere con la messa in opera del prototipo che è stato realizzato nel corso del progetto PRODIGIT. Per questa ragione, e per rendere il prototipo quanto più affidabile possibile, all’interno del campione delle sentenze del dataset non è stata fatta alcuna distinzione qualitativa ma si è proceduto alla sommarizzazione di tutte le pronunce presenti.

3. Sintesi dei criteri della massimazione civile e penale [https://www.cortedicassazione.it/resources/cms/documents/SINTESI\\_CRITERI DELLA MASSIMAZIONE CIVILE E PENALE.pdf](https://www.cortedicassazione.it/resources/cms/documents/SINTESI_CRITERI DELLA MASSIMAZIONE CIVILE E PENALE.pdf). Cfr. L. Nazzicone, *Tecniche di massimazione delle sentenze*, Roma, 2017.

4. Sono, quindi, di competenza dell’Ufficio del Massimario anche le massime in materia tributaria, poiché, come noto, la competenza per il giudizio di legittimità in questo settore è della Quinta Sezione Tributaria della Cassazione Civile.

5. Secondo il sito ufficiale della Corte di Cassazione “compito istituzionale dell’Ufficio del Massimario e del Ruolo è l’analisi sistematica della giurisprudenza di legittimità, condotta allo scopo di creare le condizioni di un’utile e diffusa informazione [...], necessaria per il miglior esercizio della funzione nomofilattica” (<https://www.cortedicassazione.it/it/massimario.page>).

6. [https://www.cortedicassazione.it/page/it/sintesi\\_dei\\_criteri\\_della\\_massimazione\\_civile\\_e\\_penale?contentId=NTZ1078](https://www.cortedicassazione.it/page/it/sintesi_dei_criteri_della_massimazione_civile_e_penale?contentId=NTZ1078).

Per quel che concerne la giurisprudenza tributaria di merito, e in particolare le pronunce delle Corti di Giustizia Tributaria di primo e secondo grado, la massimazione delle sentenze è stata svolta fino a poco tempo fa dagli Uffici Regionali del Massimario, su base volontaria<sup>7</sup>. Solo alcune sentenze, scelte da tali organi, venivano massimate, e il contenuto delle massime era spesso quello delle cc.dd. “*massime con fattispecie*”<sup>8</sup>.

I sommari automatici prodotti dalle tecniche sperimentate nel progetto PRODIGIT, invece, hanno lo scopo di sintetizzare il contenuto della sentenza, per consentire di valutare quali documenti risultanti dalla ricerca siano di interesse e meritino una lettura integrale. La sommarizzazione può definirsi dunque come il processo di condensazione di un ampio insieme di informazioni con l’obiettivo di produrre un documento più breve, simile ad un riassunto, contenente le informazioni più significative, o comunque quelle più rilevanti per il compito da svolgere. Al contempo, il sommario non è un mero riassunto dal momento che dedica maggior spazio ai principi di diritto che alle questioni di fatto. Per queste ragioni si è deciso di usare il termine “sommari” e “sommarizzazione”, piuttosto che “sintesi”, “riassunto” o “massima”.

Nell’ambito del progetto PRODIGIT abbiamo deciso di procedere alla sommarizzazione, proprio allo scopo di offrire i sommari come filtro di selezione per quando la quantità di materiale giuridico disponibile supererà la capacità umana di elaborarlo. Fornendo sintesi delle decisioni, giudici e avvocati hanno la possibilità di determinare più rapidamente se un precedente è rilevante per la questione in esame e decidere se è opportuno leggere la sentenza nella sua interezza. Inoltre, la sintesi può evidenziare i punti chiave di una lunga decisione, consentendo agli avvocati di concentrarsi su di essi.

Il lavoro parte da una ricostruzione delle tipologie di sommarizzazione in un contesto come quello italiano, fortemente influenzato dall’abitudine alla massimazione delle più rilevanti sentenze di legittimità e, in misura minore, di merito (para. 2). Presentiamo la metodologia usata nel progetto, in relazione alla scelta del dominio analizzato, la creazione dei *dataset* e i metodi per la valutazione dei risultati (para. 3). Illustriamo le due soluzioni tecniche adottate per la formulazione automatica dei sommari: l’utilizzo di tecniche di *Natural Language Processing* (NLP) per compiti specifici (cd. “*task-specific*”), utilizzate per la sommarizzazione estrattiva, e l’uso di *Large Language Model* (LLM) per la sommarizzazione astrattiva ed estrattiva (para. 4 e 5). Per ciascuno dei due metodi presenteremo anche i risultati sperimentali e le valutazioni espresse dagli esperti (para. 4.2 e 5.2). Presentiamo poi brevemente anche gli esperimenti condotti per la ricerca delle sentenze all’interno della banca dati prototipale (para. 6).

## 2 Stato dell’arte

Nel settore tributario le sentenze sono solitamente riassunte da operatori del diritto (para. 1). Talvolta, le sintesi dei casi sono messe a disposizione dalle stesse istituzioni che redigono le sentenze, come nel caso delle massime della Corte di Cassazione oppure della Corte di Giustizia. Altre volte sono proposte dalle banche dati in cui si trovano le sentenze (es. banche dati Wolters Kluwer e Giuffrè) oppure nella forma di *abstract* pubblicati da diverse riviste.

Tuttavia, com’è noto, la redazione di sintesi delle sentenze è un’attività che richiede molto tempo e, per questo, nel corso degli anni sono stati compiuti molti sforzi per automatizzare la sintesi delle sentenze<sup>9</sup>. Esistono due approcci principali per realizzare tali sintesi: estrattivo e astrattivo (Figura 1)<sup>10</sup>. Con la prima espressione si intende lo sviluppo di metodi automatici per selezionare le frasi più significative del testo in ingresso, con l’intento di combinarle in un nuovo sommario. Nella sommarizzazione estrattiva non viene apportata alcuna

7. <https://www.dgt.mef.gov.it/gt/web/guest/massimari-regionali>. Occorre specificare che l’articolo 1 della legge 130/2022 ha previsto l’abrogazione dell’art. 40 del D. Lgs. n. 545/1992, istitutivo degli Uffici dei massimari regionali, a decorrere dal 1° gennaio 2023, e l’istituzione, presso il Consiglio di Presidenza della Giustizia Tributaria, dell’Ufficio del massimario nazionale.

8. Si veda *supra* nota 1.

9. A. Deroy, K. Ghosh, S. Ghosh, *How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization?*, in ICAIL “Legal AIIA” 2023, pp. 8-19.

10. A. Nenkova, K. McKeown, *A Survey of Text Summarization Techniques*, Springer US, 2012, pp. 43-76; W. S. El-Kassas, C. R. Salama, A. A. Rafea, H. K. Mohamed, *Automatic text summarization: A comprehensive survey*, in *Expert Systems with Applications*, 2021, n. 165, pp. 113679 ss.

modifica al contenuto testuale delle frasi estratte. Questo approccio ha il vantaggio di garantire che tutto il contenuto del riassunto sia ottenuto dal documento di *input*, senza alcuna aggiunta spuria. Inoltre, esso potrebbe facilmente dare al lettore la possibilità di spostarsi dalle frasi selezionate alla loro posizione nel documento originale, in modo da ottenere un contesto per tali frasi, quando necessario. L'approccio estrattivo, però, può non riuscire a catturare tutti i contenuti rilevanti oppure può farlo solo a costo di riprodurre ampie parti dei testi originali, limitando lo scopo stesso della sintesi.

La sommarizzazione astrattiva si propone di generare un nuovo testo che fornisce una dichiarazione sinottica del contenuto dei documenti di *input*, senza riprodurne la formulazione. L'approccio astrattivo ha il vantaggio di fornire un testo breve che, in una forma linguistica appropriata, cattura il contenuto saliente di un documento molto più ampio. Esso può essere più appropriato quando il ragionamento giuridico è sviluppato in frasi lunghe, con contenuti misti, o quando una lunga decisione deve essere riassunta in un resoconto breve e chiaro, indipendentemente dal modo in cui questo contenuto è espresso nel documento originale. Inoltre, un sommario astrattivo è più simile ai riassunti svolti da essere umani che, nell'attività di sintesi, parafrasano il testo integrale piuttosto che riprodurre alcuni estratti.

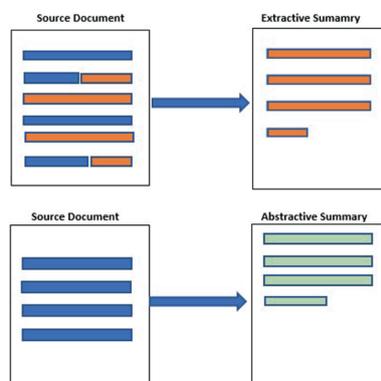


Figura 1 – Sommari estrattivi e sommari astrattivi

Nel corso del progetto abbiamo prima di tutto applicato modelli di sintesi estrattiva. La sintesi estrattiva dei testi è da tempo un ambito di ricerche dell'informatica giuridica. Nel 2004 un contributo fondamentale in questo campo è stato dato dal progetto sulle decisioni della *House of Lords Judgement (HOLJ)* che ha fornito la sintesi estrattiva di un dataset di 188 sentenze riferite al periodo 2001-2003<sup>11</sup>. Tuttavia, solo di recente i lavori in questo ambito hanno iniziato a produrre risultati promettenti, grazie all'uso delle più recenti tecniche di NLP e dell'apprendimento automatico. In PRODIGIT, abbiamo sperimentato una varietà di metodi esistenti, tra cui tecniche di ottimizzazione<sup>12</sup>, apprendimento *multi-task*<sup>13</sup>, classificazione basata sull'apprendimento automatico<sup>14</sup>, etc. I modelli estrattivi testati includono sia quelli non supervisionati<sup>15</sup> che quelli supervisionati<sup>16</sup>.

Grazie all'avanzamento nel campo del NLP, la comunità scientifica si è recentemente orientata anche verso la sintesi astrattiva o generativa per la produzione di sommari più naturali e coerenti. In particolare, sono stati

11. C. Grover, B. Hachey, I. Hughson, *The HOLJ corpus. Supporting summarisation of legal texts*, In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, pp. 47-54 (2004).

12. P. Bhattacharya et al., *Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents*, in *ICAIL 2021*, pp. 22-31.

13. A. Agarwal et al., *Extractive Summarization of Legal Decisions using Multi-task Learning and Maximal Marginal Relevance*, in *EMNLP (Findings)*, 2022, pp. 1857-1872.

14. S. Liu et al., *Combining relevance language modeling and clarity measure for extractive speech summarization*, in *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2015, pp. 957-969.

15. P. Bhattacharya et al., *Incorporating Domain Knowledge*, cit.

16. A. Agarwal et al., *Extractive Summarization*, cit.

sviluppati modelli di sintesi astrattiva per le sentenze della Corte Suprema Brasiliana<sup>17</sup> e delle Corti Supreme Inglese e Indiana<sup>18</sup>. In questi lavori viene applicata la sintesi astrattiva dei giudizi dei casi legali utilizzando l'implicazione testuale<sup>19</sup>.

Lo sviluppo di metodi di sintesi astrattiva è stato facilitato dall'utilizzo degli LLM<sup>20</sup>. Tali strumenti sono già utilizzati per la sintesi in altri domini, ad esempio per sintetizzare il contenuto di articoli di giornale<sup>21</sup>. Tuttavia, i documenti giuridici hanno alcune caratteristiche peculiari che potrebbero rendere l'applicazione di tali strumenti meno efficace nel settore della giustizia. Infatti, le sentenze sono documenti più lunghi, presentano una struttura più rigida, utilizzano molti termini tecnici e le citazioni hanno un ruolo più importante che in altri settori<sup>22</sup>. Esiste un altro lavoro, oltre a quello che si sta presentando, che utilizza gli LLM per la sintesi astrattiva in ambito giuridico<sup>23</sup>. In tale lavoro si sono applicati modelli di sintesi astrattiva e LLM preaddestrati su sentenze di tribunali indiani. Gli esiti dimostrano che i riassunti astrattivi hanno valutazioni leggermente superiori rispetto a quelli estrattivi; al contempo essi presentano frasi incomplete o prive di significato e cd. "allucinazioni" (ad esempio, errori tipici riguardano date e nomi di persone).

### 3 Metodologia

Le attività hanno seguito un processo di sviluppo a spirale dove, per ogni iterazione, abbiamo proceduto alla:

1. pianificazione delle attività sulla base di letteratura ed iterazioni precedenti;
2. prototipazione con *feedback* costante da parte degli esperti del dominio tributario nel gruppo di lavoro;
3. validazione formale di fine ciclo.

Nel processo di sviluppo abbiamo garantito l'interazione costante tra tributaristi e informatici. Da un lato, alcuni dei risultati ottenuti sono stati sottoposti agli esperti giuridici per una valutazione rapida volta a migliorare le tecniche adottate e, nel caso di utilizzo degli LLM, la formulazione del *prompt*. Dall'altro, al termine di ogni fase, i risultati ottenuti sono stati validati tramite questionari di valutazione della qualità dei sommari (para. 4.2 e 5.2).

L'adesione a tale modello ha comportato che nelle attività di sviluppo abbiamo iniziato con tecniche semplici per poi passare a soluzioni via via più articolate, secondo un modello incrementale/scalare. Tale scelta ha richiesto valutazioni rapide e periodiche.

In generale, l'attività di sviluppo incrementale si è articolata in tre fasi principali: durante la prima fase ci siamo concentrati su un gruppo ampio di modelli di sommarizzazione per individuare quelli con prestazioni migliori per il *task* di sommarizzazione, al fine di ridurre il numero di modelli da considerare. La seconda fase si è incentrata sull'insieme di modelli preferiti all'esito della prima valutazione, con l'obiettivo di verificare le

---

17. D. de Vargas Feijo, V. P. Moreira, *Improving abstractive summarization of legal rulings through textual entailment*, in *Artificial Intelligence and Law*, 2023, pp. 91–113.

18. A. Shukla, P. Bhattacharya, S. Poddar, R. Mukherjee, K. Ghosh, P. Goyal, S. Ghosh, *Legal case document summarization: Extractive and abstractive methods and their evaluation*, in *Proceedings of the Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2022, pp. 1048–1064.

19. L'implicazione testuale è una relazione ottenuta tra frammenti di testo che in qualche modo si implicano reciprocamente. L'automazione dei sistemi di riconoscimento dell'implicazione testuale è a supporto di un'ampia varietà di task basati sul testo. Si vedano D. Z. Korman, E. Mack, J. Jett, A.H. Renear, *Defining textual entailment*. *Journal of the Association for Information Science and Technology*, 2018, pp. 763 ss.

20. A. Deroy et al, *How Ready are Pre-trained Abstractive Models*, cit.

21. A. Fan, D. Grangier, M. Auli, *Controllable abstractive summarization*, in *Proceedings of the 2<sup>nd</sup> workshop on neural machine translation and generation*, 2018, pp. 45–54; Y. Liu Y, *Fine-tune BERT for extractive summarization*, 2019, in arXiv preprint arXiv: 19031 0318; T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, T. B. Hashimoto, *Benchmarking large language models for news summarization*, 2023, in arXiv preprint arXiv:2301.13848.

22. H. Turtle, *Text retrieval in the legal world*, in *Artificial Intelligence and Law*, 1995, pp. 5–54.

23. A. Deroy et al, *How Ready are Pre-trained Abstractive Models*, cit.

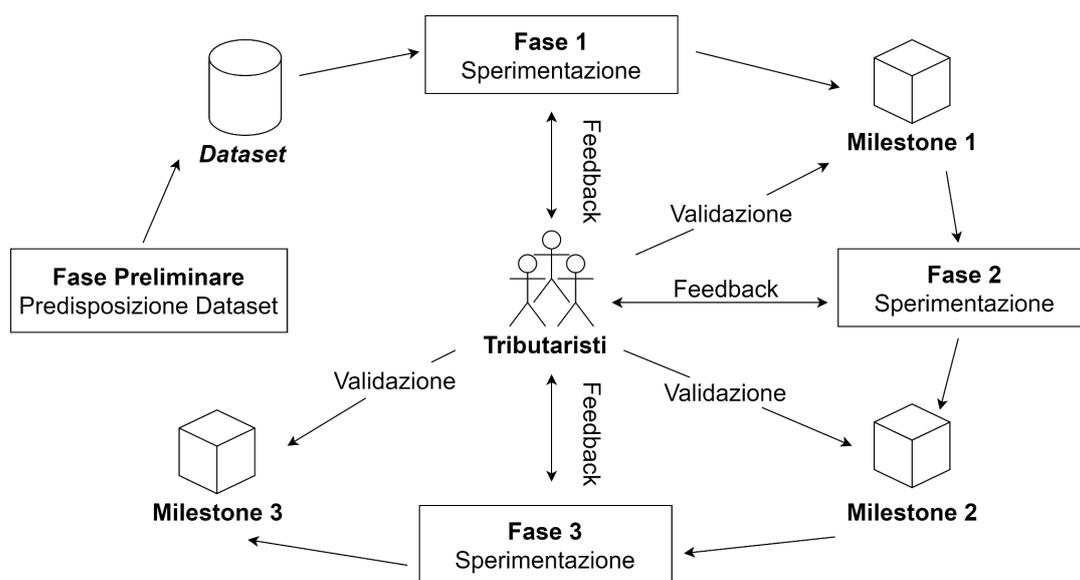


Figura 2 – Metodologia

prestazioni di questo sottoinsieme di tecnologie e scegliere quella migliore. La terza e ultima fase ha portato al perfezionamento e finalizzazione del metodo emerso come più adeguato all'obiettivo del progetto.

Per ciò che concerne le competenze, abbiamo ritenuto essenziale poter disporre di un gruppo di lavoro multidisciplinare che possedesse le competenze sia informatiche, sia giuridiche necessarie. Le prime sono state fondamentali per l'utilizzo dei sistemi di NLP e per la preparazione dei *prompt*; le seconde per poter supervisionare la creazione del *dataset* e per validare e proporre soluzioni migliorative dei risultati ottenuti dalla sommarizzazione automatica. Il primo gruppo di esperti era composto sia da membri di Società Generale d'Informatica S.p.A. (Sogei, la società di Information Technology 100% del Ministero dell'Economia e delle Finanze), sia da esperti di informatica e di informatica giuridica provenienti dall'ambito accademico. Il secondo gruppo, invece, era composto principalmente da esperti di diritto tributario provenienti dall'ambito accademico e della giustizia tributaria. Nelle attività di validazione del *dataset* finale, vale a dire quello in relazione al quale il gruppo ristretto di esperti giuridici e informatici ha ritenuto di aver ottenuto risultati migliori, le attività di validazione sono state svolte anche da esperti di diritto tributario che non erano coinvolti nelle attività di sviluppo<sup>24</sup>.

### 3.1 I dati

Come si è anticipato, l'obiettivo del lavoro è lo sviluppo di uno strumento per la sommarizzazione automatica delle sentenze di merito, per garantire agli operatori del diritto di conoscere il contenuto e le informazioni più rilevanti relative a tali sentenze senza doverne leggere l'intero contenuto. Di conseguenza, per quanto riguarda i dati, è stato necessario costruire un *dataset* di sentenze di merito che presentassero determinate caratteristiche.

Innanzitutto, abbiamo ritenuto opportuno che il *dataset* fosse costituito da sentenze dal contenuto quanto più possibile omogeneo. Infatti, le sentenze in materia tributaria spesso trattano di temi diversi: da un lato, le ragioni di ricorso o di appello riguardano sia ragioni processuali o procedurali sia ragioni di merito. Dall'altro lato, le questioni sostanziali trattate sovente si riferiscono a più imposte in ragione del principio dell'unitarietà dell'accertamento<sup>25</sup>. Tuttavia, intuitivamente, tali contenuti disomogenei potrebbero avere un impatto negativo sull'attività di sommarizzazione. Di conseguenza, non potendo escludere dal *dataset* tutte le sentenze che

24. <https://www.giustizia-tributaria.it/item.php?catid=151&itemid=4127>.

25. Il principio della tendenziale unitarietà dell'avviso di accertamento, desumibile dal D.P.R. 29 settembre 1973, n. 600, art. 43, comma 3, comporta l'inammissibilità del frazionamento dell'accertamento in caso di sopravvenuta conoscenza di nuovi elementi, sicché non

trattino anche questioni processuali o procedurali, abbiamo ritenuto opportuno concentrarci, almeno nella prima fase del lavoro, su un dominio che fosse il più omogeneo possibile dal punto di vista sostanziale.

In secondo luogo, abbiamo ritenuto che la selezione di sentenze non eccessivamente lunghe potesse agevolare l'attività di sommarizzazione automatica. Infatti, abbiamo assunto che sentenze brevi fossero in linea di massima meno complesse e che ciò potesse agevolare la sommarizzazione automatica.

In terzo luogo, abbiamo ritenuto opportuno scegliere un dominio in cui fosse disponibile un numero elevato di sentenze in un formato *machine readable*, come quelle prodotte dal sistema per la redazione di atti giudiziari PGD (Provvedimenti giurisdizionali digitali), così da rendere più agevole la trasformazione in un formato testuale. Occorre specificare che la scelta di lavorare con dati in formato PGD in ambito tributario implica che gli esperimenti possano essere svolti su sentenze recenti. Infatti, solo a partire dal 1° luglio 2019 le sentenze sono disponibili anche in tale formato.

Alla luce delle considerazioni appena esposte, abbiamo deciso di concentrarci sulle sentenze in materia di imposta di registro delle Corti di Giustizia Tributaria. Le sentenze in ambito di imposta di registro, infatti, sono generalmente più brevi di altre. Ciò dipende anche dal fatto che, nella quasi totalità dei casi, le sentenze che si occupano di tale imposta non trattano di profili relativi ad altri tributi.

Le sentenze utilizzate negli esperimenti sono adottate dalle Corti di Giustizia Tributaria di primo e secondo grado. La scelta di concentrarsi sulle sentenze sia di primo sia di secondo grado è motivata, da un lato, dalla considerazione pratica che il progetto intende sviluppare un prototipo applicabile a tutte le sentenze di merito; dall'altro di consentire la navigazione tra sentenze di primo e secondo grado. Tale scelta consente anche di verificare se le caratteristiche delle sentenze di diversi gradi di giudizio abbiano un impatto sulla qualità della sommarizzazione. Infatti, le sentenze di secondo grado spesso contengono meno riferimenti al fatto delle sentenze di primo grado e si assume che questa caratteristica potesse agevolare l'attività di sommarizzazione.

Una volta individuato il dominio delle sentenze, sono stati predisposti tre *dataset*:

1. *Dataset n. 1: 712 sentenze di 2° grado (PGD)*: Le sentenze sono state identificate da un comitato ristretto di esperti giuristi e informatici perché attinenti ad alcuni temi controversi (uso del criterio del prezzo/valore, enunciazione di atti non registrati; natura dichiarativa della divisione; identificazione degli atti a contenuto patrimoniale soggetti a registrazione in termine fisso). Le sentenze meno recenti sono del 2021, anno a partire dal quale un numero significativo di provvedimenti è reperibile in formato PGD.
2. *Dataset n. 2: 1000 sentenze di 1° grado (cartaceo e PGD)*: Nel dataset sono state incluse sia sentenze successivamente impugnate e oggetto di decisioni incluse nel Dataset n. 1, sia sentenze non appellate. La scelta di includere nel secondo *dataset* anche le sentenze di primo grado relative al Dataset n. 1 ha richiesto l'inclusione anche di sentenze non in formato PGD.
3. *Dataset n. 3: 1260 sentenze di 2° grado (PGD)*: Sentenze del 2023 su cui abbiamo scelto di concentrarci per assicurare che tutte le sentenze più recenti in materia di imposta di registro siano sommarizzate. Abbiamo ritenuto che tali sentenze potessero contenere principi più rilevanti e difficilmente già superati dalle pronunce della Corte di Cassazione.

Tutte le sentenze sono state sottoposte a un processo automatico di mascheramento e anonimizzazione, volto a schermare i riferimenti diretti e indiretti alle persone fisiche coinvolte nel contenzioso. Un *team* dedicato ha condotto a campione un'analisi della qualità dei risultati per garantire la correttezza del mascheramento.

Sia nel caso delle sentenze PGD che in quelle in formato cartaceo è stato necessario l'utilizzo di tecniche di *Optical Character Recognition* (OCR). L'OCR su PDF consente di ottenere del testo copiabile e modificabile, facilitando la gestione e la manipolazione dei documenti digitali. D'altra parte, l'estrazione di testo mediante OCR presenta una maggiore complessità dovuta alla diversità nella qualità della scansione, all'allineamento irregolare del testo e ad eventuali imperfezioni del documento cartaceo. La maggiore complessità nell'estrazione

---

è consentito all'Ufficio sostituire un primo atto di accertamento con altro successivo sulla base degli stessi elementi diversamente valutati (v. Cass., Sez. 5, Sent., 3 giugno 2015, n. 11421; Cass. civ., Sez. V, Sent., 03 giugno 2015, n. 11421; Cass., Sez. 5, Sent., 20 dicembre 2016, n. 26279).

da documenti cartacei richiede spesso una preelaborazione accurata, come la correzione dell'allineamento e la pulizia dell'immagine, al fine di ottenere risultati più precisi.

Per l'elaborazione delle sentenze di primo grado non in formato PGD è stato utilizzato il servizio *Watson Discovery*, uno strumento specifico per la trasformazione del testo mediante OCR. Si fa uso del modello linguistico GPT-4 per la correzione di eventuali errori di acquisizione.

## 3.2 I modelli

Per quanto riguarda le tecnologie, in un primo momento, ci siamo focalizzati su tecniche specifiche per la sommarizzazione estrattiva. In assenza di precedenti progetti che avessero avuto risultati soddisfacenti nella sintesi di sentenze in materia tributaria (para. 2), abbiamo ritenuto opportuno partire proprio dai sommari estrattivi (para. 4). Si è pensato, infatti, che i sommari estrattivi fossero più facilmente ottenibili allo stato dell'arte dell'evoluzione tecnologica e che fossero precisi e aderenti al testo.

In una seconda fase, abbiamo deciso di includere anche gli LLM nelle attività di sommarizzazione. Infatti, l'evoluzione delle tecniche di apprendimento automatico e dei modelli generativi, di cui GPT di OpenAI è l'esempio più noto, hanno portato a miglioramenti significativi nell'ambito dell'NLP, tra cui la sommarizzazione estrattiva<sup>26</sup>. Inoltre, l'utilizzo degli LLM consente di estrarre anche sommari astrattivi, più brevi di quelli estrattivi e più simili alle sintesi cui sono abituati gli operatori del diritto. Infatti, gli operatori giuridici in ambito tributario sono soliti interfacciarsi con sintesi o *abstract* elaborati da esperti del settore, oppure con massime della Corte di Cassazione o della giurisprudenza di merito.

### 3.2.1 Tecniche *task-specific* per la sommarizzazione estrattiva

Nell'ambito dell'estrazione automatica dai testi, strumenti di NLP *task-specific* per la sommarizzazione estrattiva sono in uso da anni e permettono di identificare ed estrarre frasi rilevanti e informazioni da documenti, articoli o altri contenuti testuali. Nell'ambito di PRODIGIT, sono stati utilizzati i seguenti strumenti:

- *Base con NLTK (NLTK)*: Algoritmo per la sommarizzazione basato sulla sola frequenza delle parole che utilizza NLTK, una libreria di Python per l'elaborazione del linguaggio naturale che fornisce strumenti e risorse per l'analisi del testo, tra cui *tokenizzazione*, *stemming*, analisi grammaticale e classificazione del testo<sup>27</sup>.
- *Luhn (LUHN)*: Metodo di estrazione del testo che identifica frasi chiave all'interno di un documento basandosi sulla frequenza delle parole e sulla loro posizione nel testo, assegnando punteggi alle frasi in base a questi criteri<sup>28</sup>.
- *Latent Semantic Analysis (LSA)*: Tecnica di riduzione dimensionale che identifica i concetti principali all'interno di un corpus testuale. Il riassunto viene generato selezionando le frasi che meglio rappresentano i concetti chiave del testo originale<sup>29</sup>.
- *TextRank (TRUNK)*: algoritmo che calcola l'importanza delle frasi in un testo, creando un grafo dove i nodi sono le frasi e i collegamenti la loro similarità. Le frasi con maggior peso, calcolato a partire dal peso delle frasi simili, sono considerate le più importanti.

26. A. Derooy, K. Ghosh, S. Ghosh, *Ensemble methods for improving extractive summarization of legal case judgements*, in *Artificial Intelligence and Law*, 2024, n. 32, p. 231 ss.

27. S. Bird, *NLTK: the natural language toolkit*, in *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions 2006*, pp. 69-72.

28. Sempre con riferimento alla sommarizzazione di testi giuridici, si veda K. Agrawal, *Legal Case Summarization: An Application for Text Summarization*, in *2020 International Conference on Computer Communication and Informatics (ICCCI)* Coimbatore, India, 2020, pp. 1-6.

29. L'analisi semantica nasce nel settore del c.d. *information retrieval*. Sul punto si veda S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, *Indexing by latent semantic analysis*, in *Journal of the American society for information science*, 1990, pp. 391-407.

- *LexRank (LEX)*: Simile a TextRank ma con un metodo per il calcolo della similarità che considera maggiormente la componente semantica e la frequenza delle parole<sup>30</sup>.

### 3.2.2 Large Language Models

Gli LLM possono essere utilizzati per una vasta gamma di compiti senza dover essere addestrati specificamente per ciascuno di essi. Ad esempio, possono completare frasi, rispondere a domande, tradurre testi ecc. Sono addestrati su ampissimi *dataset* comprendenti porzioni significative di testo scritto in linguaggio naturale (ad es. tratto da pagine web), non necessariamente relativo a un particolare dominio. L'addestramento su grandi quantità di dati consente di apprendere modelli linguistici complessi, con miliardi di parametri. Questo permette loro di catturare una vasta gamma di *pattern* linguistici e conoscenze semantiche.

Tra gli LLM, sono stati sperimentati i seguenti modelli linguistici:

- IT5: prima famiglia di modelli addestrati specificatamente per la comprensione e la generazione della lingua italiana<sup>31</sup>. Sono state utilizzate le varianti *it5-small*<sup>32</sup> e *it5-large*<sup>33</sup>, che si differenziano principalmente per dimensione (numero di parametri) e numero di livelli.
- GPT (*Generative Pre-trained Transformer*): modello sviluppato da OpenAI e attualmente considerato il più efficiente tra i modelli disponibili. GPT è stato utilizzato sia nella versione 3.5 (*text-davinci-003*)<sup>34</sup> che nella versione 4 (*gpt-4-32k*)<sup>35</sup>. Quest'ultima è la versione più recente e, secondo i suoi sviluppatori, più capace della precedente in compiti come la *sentiment analysis* e la classificazione dei testi. GPT-4 è inoltre in grado di elaborare un numero più significativo di *token* in ingresso e in uscita (circa 32 mila), consentendo così di effettuare compiti più sofisticati.

La famiglia di modelli IT5 segna un notevole progresso nel campo dei modelli linguistici per l'italiano. Abbiamo deliberatamente scelto di concentrarci su questo modello specifico con attenzione. Infatti, sebbene modelli "generali" come GPT abbiano dimostrato buone prestazioni in diverse lingue, compreso l'italiano, abbiamo ritenuto che un modello specifico potesse raggiungere un livello di precisione e fluidità superiore, adattandosi meglio alle peculiarità della nostra lingua.

In ogni caso, abbiamo deciso di utilizzare anche GPT negli esperimenti. Infatti, i modelli appartenenti alla famiglia GPT hanno dimostrato in varie occasioni di essere attualmente i più performanti in molti *task* di comprensione e generazione del testo, e quindi di essere un valido riferimento per l'intera categoria.

### 3.3 La validazione

Il metodo adottato ha richiesto che si individuassero delle tecniche e dei criteri di validazione. Abbiamo considerato la possibilità di automatizzare il processo di valutazione dei sommari. Tuttavia, gli strumenti automatici esistenti per la valutazione della qualità dei sommari non si prestano ad essere applicati in modo utile alla sintesi astrattiva. Infatti, qui la valutazione dovrebbe tenere conto di aspetti quali la necessità di comprendere il contesto e la difficoltà nel misurare aspetti come la coerenza semantica e l'efficacia comunicativa<sup>36</sup>. Inoltre, essa deve considerare come vengono gestite informazioni omesse, aggiunte o riformulate nei riassunti. Tutto ciò rende l'automatizzazione del processo di valutazione particolarmente complessa.

---

30. G. Erkan, D. Radev *Lexrank: Graph-based lexical centrality as salience in text summarization*, in *Journal of artificial intelligence research*, 2004, pp. 457-479.

31. G. Sarti, M. Nissim, *It5: Large-scale text-to-text pretraining for italian language understanding and generation*, in *arXiv preprint arXiv:2203.03759* 2022.

32. <https://huggingface.co/gsarti/it5-small>.

33. <https://huggingface.co/gsarti/it5-large>.

34. T.B. Brown et al, *Language Models are Few-Shot Learners*, 2020, in *NeurIPS*, 2020.

35. OpenAI: GPT-4 Technical Report, 2023.

36. Si veda sul punto E. Lloret, L. Plaza, A. Aker, *The challenging task of summary evaluation: an overview*, in *Lang Resources & Evaluation*, 2018, pp. 101-148.

Per questo motivo, abbiamo optato per una valutazione da parte di esperti di diritto tributario attraverso questionari che permettessero di valutare gli esiti delle sintesi estrattive e astrattive. I questionari sono stati precedentemente sottoposti al comitato etico del progetto PRODIGIT, che li ha esaminati, ha proposto perfezionamenti e chiarimenti sulle domande e sulla metodologia.

### 3.3.1 La procedura di valutazione

Fase	N. esperti	Tipologie di sommario	Aspetti valutati	Dati	Blind
1	13	7 estrattivi	Correttezza, Forma	Dataset n. 1	Sì
		4 astrattivi	Completezza, Soddisfazione		
Questionari (3 decisioni per tipologia sommario per valutatore)					
2	43	3 astrattivi	Correttezza, Forma	Dataset n. 1	Sì
			Completezza, Soddisfazione		
Questionari (5 sentenze per tipologia sommario per valutatore)					
3	43	1 astrattivo	-	Dataset n. 1 e 3 (1524 sentenze)	-
		Correzione sommari (~35 sommari per valutatore)			

Figura 3 – Le fasi della valutazione

L'attività di validazione ha avuto luogo in tre fasi principali, aggiuntive rispetto alla validazione informale in corso d'opera (Figura 3).

La *prima fase* ha coinvolto 13 esperti che hanno validato tecniche sia estrattive sia astrattive. In particolare, sono state validate 7 tecniche estrattive e 4 tecniche astrattive. Le 11 tecniche sono state applicate ai sommari relativi ad alcune sentenze appartenenti al Dataset n. 1, e 12 esperti hanno validato i sommari creati con tutte le tecniche su 3 delle decisioni appartenenti all'insieme. Ciascun valutatore ha esaminato le 11 tipologie di sommario attraverso questionari distinti per sommari estrattivi e astrattivi, senza che gli venissero fornite indicazioni sulla tecnica utilizzata.

La *seconda fase*, che si assumeva avrebbe condotto alla scelta della tecnica di sommarizzazione migliore, ha coinvolto un numero superiore di valutatori. Gli esperti, infatti, erano 43. Dopo aver constatato la preferenza dei valutatori per i sommari astrattivi, sono stati testati solo 2 modelli (GPT-4 e IT5). IT5 è stato utilizzato per la sommarizzazione astrattiva a testo libero, mentre GPT-4 per la sommarizzazione astrattiva a tema e a testo libero (para. 4.1 e 5.1). Dunque, in totale sono state validate tre tipologie di sommarizzazione (IT5 a testo libero; GPT-4 a testo libero; GPT-4 a tema). Si è scelto di non testare nuovamente GPT-3 dal momento che era simile a GPT-4. Ciascun valutatore ha esaminato i sommari prodotti dai 3 modelli in relazione a 5 decisioni estratte dal Dataset n. 1, per un totale di 15 sommari. Ancora una volta, i valutatori non erano a conoscenza della tecnica utilizzata.

La *terza fase* si è concentrata sulla validazione di GPT-4, valutato più positivamente di IT5 nella seconda fase. Gli esperti coinvolti sono stati 40 (i c.d. "*writers*"). In questo caso la validazione ha riguardato GPT-4 usato per la sommarizzazione combinata, in grado di fornire risultati più simili a quelli standard, combinando i *prompt* per la sommarizzazione a testo libero e a tema (para. 5.2). I valutatori hanno valutato 1524 sommari generati dal Dataset n. 3 e dal Dataset n. 1. Infatti, abbiamo sommarizzato tutte le sentenze del Dataset n. 3 e, per avere un numero maggiore di dati, abbiamo incluso anche le sentenze più recenti (quelle relative al 2023) del Dataset n. 1.

### 3.3.2 I criteri di valutazione

Nella prima e seconda fase di valutazione, i questionari richiedevano agli esperti del dominio di valutare la loro soddisfazione in relazione a diversi parametri: la correttezza, la forma, la completezza e la soddisfazione complessiva.

Per *correttezza* si intende l'accuratezza nel cogliere i punti chiave, le questioni giuridiche e le informazioni essenziali dei documenti di partenza. Questo criterio risponde alla basilare esigenza che la sintesi corrisponda al contenuto della sentenza originaria.

Per *forma* si intendono la coerenza, la leggibilità, la correttezza sintattica e grammaticale e l'aderenza al lessico giuridico. Questo criterio risponde all'esigenza di bilanciare la comprensibilità del riassunto con l'aderenza al linguaggio tecnico giuridico. Infatti, non solo il sommario deve essere corretto ma deve essere comprensibile e utilizzare un linguaggio intellegibile per il giurista perché, da un lato, il diritto è un linguaggio tecnico e dunque è fondamentale che i sommari utilizzino in modo appropriato termini e concetti giuridici; dall'altro lato, se così non fosse, il sommario potrebbe essere percepito come non affidabile dal lettore.

Per *completezza* si intende l'inclusione di tutti gli elementi rilevanti della decisione e la rappresentazione completa del contenuto della fonte. Infatti, l'utilità di un sommario risiede proprio nella capacità di fornire all'interprete il contenuto essenziale della decisione in poche righe, senza però omettere elementi rilevanti. Tale caratteristica è fondamentale quando il sommario viene utilizzato dal giurista per decidere se procedere o meno alla lettura integrale del testo della sentenza.

Infine, la *soddisfazione complessiva* è il grado di soddisfazione per la qualità globale della sintesi. L'obiettivo è di valutare ulteriori aspetti del sommario non considerati dai precedenti criteri che possono influire sulla soddisfazione complessiva dell'operatore, ad esempio l'eccessiva lunghezza del sommario o l'inclusione di informazioni irrilevanti.

Gli esperti di dominio hanno letto un certo numero di sentenze e i relativi sommari generati automaticamente. Dopodiché hanno compilato il questionario attribuendo a ciascun criterio un punteggio da 1 a 5. Inoltre, il questionario permetteva agli esperti di fornire una valutazione qualitativa ulteriore rispetto ai criteri indicati inserendo commenti a testo libero.

Per garantire l'affidabilità delle valutazioni abbiamo deciso che i revisori non avrebbero saputo con quali modelli erano state generate le sintesi. La valutazione complessiva dei sommari è stata ottenuta sommando i voti espressi dagli esperti e attribuendo ad ogni modello gli ulteriori eventuali commenti qualitativi inseriti dai valutatori. Attraverso il processo di valutazione abbiamo individuato i punti di forza e di debolezza per ciascun modello e i possibili profili di miglioramento dei modelli preferiti.

Per la terza e ultima valutazione, abbiamo scelto di mantenere la *soddisfazione complessiva* come unica metrica. Questa scelta è stata motivata dal fatto che, pur riducendo i parametri, i valutatori potevano intervenire direttamente sul sommario per proporre modifiche formali, sintattiche, concettuali e di contenuto. Per garantire la correttezza e l'uniformità delle correzioni sono state svolte sessioni di formazione e monitoraggio in corso d'opera. Inoltre, era richiesto ai valutatori di motivare le ragioni della modifica. I risultati di questa terza analisi sono stati elaborati da Sogei per calcolare il grado di scostamento tra il sommario generato automaticamente e quello corretto. Questo metodo di validazione, basato su modifiche specifiche piuttosto che sulla valutazione in base a parametri, ha consentito di verificare dal punto di vista qualitativo gli errori dei sommari generati automaticamente.

## 4 Sommarizzazione estrattiva

Per meglio comprendere i diversi metodi di sommarizzazione testati riportiamo di seguito alcuni esempi riferiti sia alla sommarizzazione estrattiva (para. 4), sia a quella astrattiva (para. 5). Tutti gli esempi sotto riportati si riferiscono alla medesima sentenza, emanata dalla Corte di Giustizia Tributaria di secondo grado della Sicilia, n. 7683 del 2022, relativa all'applicazione delle c.d. "*agevolazioni prima casa*" in materia di imposta di registro. La sentenza contiene un'unica questione giuridica rilevante (quella relativa all'applicabilità dell'agevolazione nel caso di specie) e, al contempo, una motivazione della decisione piuttosto articolata in cui sono contestualmente presi in considerazione elementi di fatto (la dimensione dell'immobile e il suo stato di conservazione) e di diritto (la corretta interpretazione della norma agevolativa)<sup>37</sup>.

---

37. La questione sottoposta ai giudici riguarda la possibilità (come sostenuto dal contribuente e dal giudice di prime cure) o meno (come invece sostenuto dall'Amministrazione) di riconoscere le agevolazioni prima casa (imposta di registro ridotta dal 9% al 2%) ad un soggetto già proprietario di un immobile, pur inadeguato per dimensioni ad accogliere l'intero nucleo familiare. Il giudice di appello disconosce le pretese dell'Amministrazione, secondo cui le agevolazioni non dovrebbero essere riconosciute, sulla base di un'interpretazione sistematica e non letterale delle disposizioni agevolative, già affermata dalla Corte di Cassazione.

## 4.1 Descrizione e risultati

La sommarizzazione estrattiva, su cui si è concentrata la prima fase del lavoro, è avvenuta usando sia tecnologie di NLP *task-specific*, sia LLM per l'estrazione di sommari estrattivi. In un primo momento, abbiamo applicato gli strumenti NLP *task-specific* alle decisioni giudiziarie per ottenere sintesi da sottoporre alla valutazione di esperti. Con tali metodi di NLP sono stati ottenuti riassunti particolarmente lunghi, come evidenziato nell'esempio riportato.

Anche dopo tale innovazione legislativa e, dunque, in relazione al testo vigente, la giurisprudenza prevalente del supremo collegio (vedi da ultimo cass. civ n 20981/2021) ha aderito all'opzione interpretativa secondo cui la mera titolarità di un cespite immobiliare non è ostativa al riconoscimento dell'agevolazione, la quale spetta, invece, al contribuente impossidente di un immobile che possa essere adibito ad abitazione ( in tal senso anche cass., sez. 5, ordinanza n. 19989 del 27/07/2018, secondo cui «in tema di agevolazioni fiscali per la prima casa, ai sensi dell'art. 1, nota ii bis, della tariffa allegata al d.p.r. n. 131 del 1986, nel testo (applicabile "ratione temporis") modificato dall'art. 3, comma 131, della l. n. 549 del 1995, il concetto di "idoneità" della casa pre-posseduta - ostativo alla fruizione del beneficio (ed espressamente previsto nella previgente normativa) - deve ritenersi intrinseco alla nozione stessa di "casa di abitazione", da intendersi quale alloggio concretamente idoneo, sia sotto il profilo oggettivo-materiale che giuridico, a soddisfare le esigenze abitative dell'interessato»; nonché cass., sez. 5, sentenza n. 2565 del 02/02/2018, che ha statuito che «in tema di agevolazioni prima casa ... «l'idoneità» dell'abitazione pre-posseduta va valutata sia sotto il profilo oggettivo- effettiva inabitabilità-, che sotto quello soggettivo - fabbricato inadeguato per dimensioni o caratteristiche qualitative- , nel senso che il beneficio trova applicazione anche nell'ipotesi di disponibilità di un alloggio che non sia concretamente idoneo, per dimensioni e caratteristiche complessive, a soddisfare le esigenze abitative dell'interessato.» e nello stesso senso si sono altresì pronunciate anche cass., sez. 6-5, ordinanza n. 5051 del 24/02/2021, cass., sez. 6-5, ordinanza n. 18091 del 05/07/2019, e cass., sez. 6-5, ordinanza n. 18092 del 05/07/2019).

Oltre ai metodi NLP *task-specific*, abbiamo utilizzato anche gli LLM, sebbene essi non siano naturalmente destinati a compiti estrattivi. Infatti, data la generale scarsa soddisfazione degli esperti riguardo all'estrazione per mezzo di NLP (para. 4.2), si voleva verificare se gli LLM dessero risultati migliori.

A tal fine, abbiamo creato un *prompt* per permettere agli LLM l'estrazione. Abbiamo effettuato tale estrazione<sup>38</sup> con GPT-3, GPT-4, e IT5, anche se quest'ultimo ha mostrato fin da subito risultati non soddisfacenti. Infatti, il modello IT5 non riusciva a seguire le istruzioni per la redazione della sintesi e riprodurre il testo originale senza modifiche, arrivando sempre a parafrasarlo.

Il *prompt* utilizzato per i sommari estrattivi è diviso in tre sezioni.

Nella *prima sezione* (Figura 4, riga 2,) si definisce il *task* di "sommarizzazione estrattiva" inteso come un testo in *output* composto da un sottoinsieme delle frasi di un testo in *input*.

Nella *seconda sezione* (Figura 4, riga 3) si descrive il metodo per ottenere un riassunto: si suggerisce cioè al modello di dividere il documento in frasi, assegnare un punteggio a ciascuna frase in base alla sua rilevanza semantica e sulla base dei punteggi identificare le frasi più significative.

Nella *terza sezione* (Figura 4, riga 5) è indicato il formato in cui deve essere prodotto il testo in *output*, cioè come elenco di frasi e punteggio assegnato alla loro importanza semantica. L'assegnazione del punteggio ha reso più efficace l'elaborazione del sommario sulla base del *prompt*.

La sintesi estrattiva ottenuta con gli LLM era nella gran maggioranza dei casi più breve di quella fornita dagli strumenti di NLP *task-specific*. Nel caso della sentenza assunta a esempio, il sommario in *output* usando GPT-4 era il seguente:

---

38. Generazione eseguita con *temperatura* a 1.0 e nessun vincolo sulla lunghezza del sommario. I restanti parametri sono stati mantenuti al loro valore predefinito (<https://platform.openai.com/docs/api-reference/completions/create>), ([https://huggingface.co/docs/transformers/main\\_classes/text\\_generation](https://huggingface.co/docs/transformers/main_classes/text_generation)).

```

1 Fra graffe è presente il contenuto di una sentenza.
2 Voglio costruire un riassunto estrattivo, cioè un testo
  composto da un sottoinsieme delle frasi presenti nel testo
  in input.
3 Per fare un riassunto estrattivo dovrei prima suddividere
  il testo in frasi, valutare l'importanza semantica,
  classificarle in base a questa importanza e selezionare le
  K frasi più rilevanti.
4 In questo caso, K è uguale a 5.
5 Le frasi devono essere presentate nello stesso ordine in
  cui appaiono nel testo originale secondo la seguente
  struttura.
6 [1] Frase 1 [Punteggio 1]
7 7 ...
8 [K] Frase 1 [Punteggio K]
9 9
10 { ... }

```

Figura 4 – Il\* prompt \*per la sommarizzazione estrattiva

Questione preliminare fondamentale è quella concernente la valutazione della idoneità o meno dell'immobile pre-acquistato dal ricorrente, rispetto alle esigenze abitative proprie e della propria famiglia. Risulta dalla stessa che l'immobile di che trattasi è composto da un piccolo vano e un bagno (raffigurati nella allegata documentazione fotografica), avente una superficie catastale complessiva di mq. 25,00 ed una superficie utile abitabile di mq. 17,90 (come da planimetria allegata). Il tecnico ha altresì accertato che l'immobile presenta tracce di umidità di risalita, muffe e funghi, concludendo che esso non risulta adatto ad essere utilizzato per soddisfare i più elementari bisogni abitativi per un nucleo familiare composto da padre e due figli in età scolare. Sulla base del citato accertamento, suffragato da planimetria e fotografie, deve perciò ritenersi assodato che l'immobile acquistato dal resistente\_1 il 19/3/2003, debba essere considerato inidoneo a soddisfare le esigenze abitative del predetto e della di lui famiglia, sicché correttamente il primo giudice ha ritenuto tale acquisto non preclusivo rispetto ai benefici fiscali invocati. Pertanto, l'appello proposto va rigettato e va confermata la sentenza impugnata.

## 4.2 Validazione

La Figura 5 mostra i risultati dei questionari per i modelli di NLP *task-specific*. Si può notare come tali modelli abbiano fornito prestazioni valutate in modo simile. In particolare, la correttezza è abbastanza elevata per tutti. Ciò era prevedibile dato che i sommari estrattivi sono costituiti da frasi estrapolate dal testo originale, che si presume sia corretto linguisticamente. Al contrario, la completezza è stata valutata in modo negativo, poiché di frequente frasi significative sono state omesse. Per quanto riguarda la forma, le valutazioni piuttosto negative sono spesso dovute alla struttura sintattica dei dati in *input* (es. frasi molto lunghe, sintassi non sempre corretta) che inficia la qualità dei sommari in *output*. La soddisfazione dei valutatori rispetto ai sommari estrattivi è complessivamente bassa. Ciò è stato evidenziato anche dai commenti a testo libero degli esperti, che sottolineano l'esigenza di abbreviare i sommari.

La Figura 6 mostra i risultati di GPT-3 e GPT-4 nella sommarizzazione estrattiva. Si nota un miglioramento rispetto alla correttezza dei sommari ma un risultato piuttosto basso in termini di completezza. Anche in questo caso frasi significative non risultano incluse nel sommario, specialmente quando si tratta di decisioni lunghe. Le valutazioni negative sulla forma, malgrado si tratti di sommari astrattivi, sono dovute al fatto che GPT-4 spezza le frasi eliminandone alcune parti.

	LEX	LSA	TRUNK	LUHN	NTLK
Forma	2.85 (1.23)	3.00 (1.11)	3.09 (1.00)	2.73 (0.96)	3.30 (0.90)
Completezza	3.15 (1.51)	2.85 (1.29)	2.82 (1.27)	2.45 (0.99)	2.40 (1.43)
Correttezza	3.69 (1.26)	3.54 (1.39)	3.91 (1.00)	3.73 (1.14)	3.50 (1.20)
Soddisfazione	2.69 (1.32)	2.77 (1.19)	2.45 (1.08)	2.36 (0.88)	2.50 (1.36)

Figura 5 – Sommarizzazione estrattiva con tecniche di NLP task-specific

	GPT4	GPT3
Forma	3.10 (0.54)	2.80 (0.75)
Completezza	3.00 (0.89)	3.00 (1.10)

Figura 6 – Sommarizzazione estrattiva basata su LLM

Vista la valutazione complessivamente insoddisfacente dei metodi estrattivi, indipendente dall'uso di modelli specifici o LLM, ci siamo concentrati sugli approcci di sommarizzazione astrattiva.

## 5 Sommarizzazione astrattiva

Come si è già detto sopra, la sommarizzazione astrattiva si propone di generare un nuovo testo consistente in una sintesi del documento di *input*, senza riprodurne la formulazione come nella sommarizzazione estrattiva.

### 5.1 Descrizione e risultati

In particolare, in PRODIGIT, abbiamo tentato di generare tre tipologie di sommari astrattivi: (i) sommari astrattivi “a testo libero”; (ii) sommari astrattivi “a tema”; (iii) sommari che combinano i precedenti tipi. Quanto alle tecniche, abbiamo sperimentato IT5 ed entrambe le versioni di GPT (GPT 3.5. e GPT 4)<sup>39</sup>.

#### 5.1.1 Sommarizzazione a testo libero

I sommari astrattivi a testo libero non hanno una struttura prestabilita, cioè contengono qualsiasi informazione il modello generativo ritiene rilevante sulla base di un *prompt* generico (Figura 7). Sono stati testati IT5, GPT-3 e GPT-4 per generare riassunti, in particolare, per le due sezioni delle decisioni: “svolgimento del procedimento” e “motivi della decisione”. A questo scopo, abbiamo scelto di iniziare con un *prompt* molto semplice per tutti i modelli da raffinare progressivamente sulla base delle attività di valutazione.

```
1 Fai un sommario del seguente testo tra parentesi graffe
2
3 {...}
```

Figura 7 – Il prompt per la sommarizzazione astrattiva di base

Di seguito è riportata la sintesi generata per la decisione nella sentenza esempio (sezione “sviluppo del procedimento”) utilizzando GPT-4.

39. Generazione finale eseguita con *temperatura* a 0.4 e lunghezza massima del testo generato di 2500 token, se supportato dal modello. I restanti parametri sono stati mantenuti al loro valore predefinito (<https://platform.openai.com/docs/api-reference/completions/create>), ([https://huggingface.co/docs/transformers/main\\_classes/text\\_generation](https://huggingface.co/docs/transformers/main_classes/text_generation)).

Il testo riguarda la questione dell'idoneità di un immobile pre-acquistato dal ricorrente per soddisfare le esigenze abitative della sua famiglia. La perizia giurata depositata dall'appellato evidenzia che l'immobile, di piccole dimensioni e con tracce di umidità, muffe e funghi, non è adatto ad ospitare un nucleo familiare composto da padre e due figli in età scolare. Pertanto, il giudice ha ritenuto corretto non considerare tale acquisto come preclusivo per i benefici fiscali invocati. La giurisprudenza prevalente sostiene che la mera titolarità di un immobile non è sufficiente per negare l'agevolazione, che spetta invece a chi non possiede un immobile idoneo ad abitazione. L'appello proposto viene rigettato e la sentenza impugnata viene confermata, con le spese di questa fase poste a carico dell'ufficio.

I modelli GPT forniscono un risultato comparativamente migliore rispetto ai modelli IT5, sia per quanto riguarda la leggibilità, sia per la completezza. Tuttavia, è possibile notare qualche difetto, riscontrato dagli esperti: ad esempio una certa ripetitività e a volte la circostanza che la soluzione in fatto venisse confusa con il principio di diritto. Sulla base di questi difetti, abbiamo sviluppato approcci più raffinati di sommarizzazione astrattiva, specificando il relativo *prompt*.

### 5.1.2 Sommarizzazione a tema

La sommarizzazione a tema fornisce sintesi basate su questioni specifiche affrontate nel testo della sentenza. L'obiettivo è quello di distinguere le differenti questioni affrontate dai giudici e fornire un'analisi sintetica per ciascuna di esse. Tale approccio è particolarmente utile per determinare la rilevanza giuridica dei casi permettendo all'interprete di identificare ed esaminare gli elementi salienti della decisione.

Anche in questo caso, abbiamo preferito l'utilizzo di GPT-3.5 e GPT-4, dato che con IT5, l'esperimento non ha avuto successo.

Il *prompt* utilizzato per la sommarizzazione a tema fornisce indicazioni formali e contenutistiche sul testo da generare in *output*. Le indicazioni formali consistono nell'elaborazione di una sequenza di domande/risposte. Le domande sono indicate come QD1, QD2, ..., QDn, mentre le risposte sono indicate come PD1, PD2, ..., PDn. Le indicazioni contenutistiche richiedono che ciascuna risposta alla relativa domanda contenga la specificazione di un principio decisionale, ovvero un'affermazione riguardante l'applicazione o l'interpretazione di una norma esplicita o di una decisione precedente.

Dopo aver definito i principi, il *prompt* continua affermando che una domanda è un'entità linguistica a cui occorre rispondere attraverso un principio. Per assicurare che il modello si concentri su questioni essenziali ma indipendenti, sono state aggiunte alcune prescrizioni: due principi devono essere molto diversi l'uno dall'altro; il numero di principi in un testo varia tra 1 e 2; nei testi lunghi possono comparire più di due principi. È stato inoltre specificato che i principi devono essere riportati esplicitamente e che le domande devono essere formulate in termini generali, cioè senza fare riferimento al caso specifico. Abbiamo voluto evidenziare così il fondamento giuridico delle soluzioni di volta in volta adottate dal giudice in relazione alle questioni di diritto trattate, e non quello di mostrare come tali questioni trovino soluzione nel caso di specie. Il *prompt* esteso è riportato nella Figura 8:

Di seguito è riportato uno dei principi estratti utilizzando GPT-4.

**QD2:** Qual è l'interpretazione corrente della normativa sulle agevolazioni fiscali per la prima casa in relazione all'idoneità dell'abitazione pre-posseduta?

**PD2:** L'idoneità dell'abitazione pre-posseduta va valutata sia sotto il profilo oggettivo (effettiva inabitabilità) che sotto quello soggettivo (fabbricato inadeguato per dimensioni o caratteristiche qualitative), nel senso che il beneficio trova applicazione anche nell'ipotesi di disponibilità di un alloggio che non sia concretamente idoneo, per dimensioni e caratteristiche complessive, a soddisfare le esigenze abitative dell'interessato (Cass., sez. 5, ordinanza n. 19989 del 27/07/2018, Cass., sez. 5, sentenza n. 2565 del 02/02/2018).

Abbiamo sperimentato anche un *prompt* più articolato includendo istruzioni per l'identificazione della porzione di testo originale che tratta il tema riassunto (base testuale, BT) per l'estrazione di parole chiave

1 QD1: testo  
2 PD1: testo  
3  
4 QD2: testo  
5 PD2: testo  
6  
7 ...  
8  
9 QDn: testo  
10 PDn: testo  
11  
12 i principi di diritto (PD) e le questioni diritto (QD).  
13  
14 Le QD sono le domande a cui i PD rispondono.  
15 Le QD non contengono nessun riferimento al caso di specie  
e agli attori della vicenda.  
16  
17 I PD sono le interpretazioni di una o più norme contenute  
nel testo tra parentesi graffe.  
18 Per ogni PD specifica i riferimenti alle norme.  
19 Il numero di PD in un testo è di solito 1 o 2.  
20 I PD non contengono nessun riferimento al caso di specie e  
agli attori della vicenda.  
21 Due PD devono essere molto diversi tra di loro.  
22 In testi lunghi il numero di PD può essere maggiore di 2.  
23  
24 {...}

Figura 8 – Il prompt per la sommarizzazione astrattiva a tema

(KW) da tale porzione di testo. La riga 14 del *prompt* originale (Figura 8) è stata quindi sostituita dal testo riportato nella Figura 9.

```
1 I principi di diritto (PD), le questioni diritto (QD), le  
parole chiave (KW) e le basi testuali (BT).
```

Figura 9 – Precisazioni nel prompt per la sommarizzazione a tema

Mentre la riga 25 è stata ampliata come si vede nella Figura 10:

```
1 Le BT sono le porzioni del testo fra parentesi graffe più  
rilevanti per l'estrazione di un PD e di una QD.  
2 Le BT non devono contenere variazioni rispetto al testo fra  
parentesi graffe.  
3 Per ogni QD e PD, restituisci massimo tre BT.  
4  
5 Le KW identificano i temi fondamentali del testo fra parentesi  
graffe, cioè i concetti giuridici impiegati, gli oggetti  
disciplinati e le materie trattate.
```

Figura 10 – Precisazioni nel prompt per la sommarizzazione a tema

Una parola chiave si riferisce a concetti e termini, giuridici e non giuridici, rilevanti per individuare il contenuto e gli aspetti salienti del testo al quale si riferiscono. La parte del *prompt* relativa alle parole chiave segue l'estrazione dei principi giuridici, che dovrebbero fornire informazioni contestuali per restringere la generazione alle parole chiave più rilevanti con un significato giuridico sostanziale.

Il *prompt* include anche istruzioni per collegare i principi estratti alla base testuale (BT), cioè alla parte pertinente delle decisioni originali da cui sono stati estratti. Tale riferimento ci permette di verificare facilmente la corrispondenza tra i principi estratti e il testo originale. Al modello viene richiesto di estrarre al massimo tre frammenti, senza introdurre alcuna variazione rispetto al testo originale.

Di seguito sono riportati il testo e le parole chiave associate ai principi dell'esempio precedente:

**BT1:** [in tema di agevolazioni fiscali per la prima casa, ai sensi dell'art. 1, nota ii bis, della tariffa allegata al d.p.r. n. 131 del 1986, nel testo (applicabile "ratione temporis") modificato dall'art. 3, comma 131, della i. n. 549 del 1995]

**BT2:** [il concetto di "idoneità" della casa pre-posseduta - ostativo alla fruizione del beneficio (ed espressamente previsto nella previgente normativa) - deve ritenersi intrinseco alla nozione stessa di "casa di abitazione", da intendersi quale alloggio concretamente idoneo, sia sotto il profilo oggettivo-materiale che giuridico, a soddisfare le esigenze abitative dell'interessato]

**BT3:** [le agevolazioni in esame rispondono alla ragionevole ratio di favorire l'acquisto di un'abitazione nel luogo di residenza o di lavoro a vantaggio di chi, nello stesso luogo, non abbia la possidenza di un'altra casa di abitazione obiettivamente idonea a soddisfare le sue esigenze]

**KW:** [agevolazioni fiscali, prima casa, idoneità abitativa, esigenze abitative, inabitabilità, inadeguatezza, titolarità immobiliare, normativa, giurisprudenza]

### 5.1.3 Sommarizzazione combinata

Nell'ultimo tipo di sommarizzazione astrattiva, abbiamo combinato aspetti degli approcci precedenti: da un lato, la capacità del primo metodo di generare un riassunto integrato e non frammentato; dall'altro, la capacità del secondo di identificare gli aspetti giuridici più rilevanti della sentenza in *input*.

Per tali ragioni, con la sommarizzazione astrattiva combinata, ci si è posti l'obiettivo di fornire un sommario unitario, che non si presentasse come una mera serie di risposte a domande o presentazione di punti salienti ma che, al contempo, includesse sempre alcuni elementi specifici individuati nel *prompt*. In tal modo il sommario avrebbe avuto il duplice vantaggio di essere facilmente leggibile in un'unica soluzione e non omettere elementi rilevanti, presentandoli in modo sufficientemente specifico. Inoltre, avrebbe consentito di superare i limiti relativi agli altri metodi di sommarizzazione.

Questo metodo combinato implica, perciò, due fasi: i) una analitica e ii) una sintetica. La fase analitica permette di individuare specifiche informazioni chiave nella sentenza ed estrarle automaticamente attraverso l'utilizzo di *prompt* specifici. In questa fase, nel *prompt* si richiede di individuare ed estrarre le seguenti informazioni:

1. **Identificazione della Corte** (Figura 11): il modello deve identificare gli estremi della Corte che ha emesso la sentenza da sommarizzare (es. Corte di Giustizia Tributaria di secondo grado dell'Emilia-Romagna).
2. **Identificazione delle parti e delle domande** (Figura 12): il modello deve indicare le parti coinvolte nel caso, distinguendo tra ricorrente e resistente, o tra appellante e appellato. Per ciascuna parte coinvolta, deve individuare inoltre le ragioni del ricorso o dell'appello.
3. **Esito del primo grado** (Figura 13): il modello deve sinteticamente indicare l'esito del processo di primo grado (es. "La CTP di Modena ha rigettato il ricorso e regolato le spese di lite sulla base della soccombenza").
4. **Specificazione dei fatti rilevanti** (Figura 14): il modello deve individuare e descrivere accuratamente i fatti più rilevanti all'origine della controversia. L'individuazione dei fatti è utile per comprendere le questioni trattate nel caso e le decisioni della Corte.
5. **Indicazione della decisione della Corte** (Figura 15): il modello deve indicare la decisione finale della Corte e se le domande delle parti sono state accolte o rigettate.
6. **Individuazione dei principi di diritto o criteri decisionali**: il modello deve identificare i principi di diritto (se presenti), le leggi, i precedenti che la Corte ha utilizzato nella motivazione per giungere a un certo esito e le premesse in diritto (criteri decisionali) su cui si basa la motivazione della Corte.

```

1  Ti è stato fornito il testo di una sentenza di secondo
   grado in ambito tributario.
2  Chi ha emesso la sentenza?
3  Fornisci le porzioni della sentenza originale da cui hai
   estratto l'informazione senza apportare modifiche al
   testo.
```

Figura 11 – Prompt per l'identificazione della Corte

```

1  Ti è stato fornito il testo di una sentenza di secondo
   grado in ambito tributario.
2  Quali sono le parti coinvolte nell'appello e le domande
   da loro proposte?
3  Fornisci le porzioni della sentenza originale da cui hai
   estratto l'informazione senza apportare modifiche al
   testo.
```

Figura 12 – Prompt per l'identificazione delle parti e delle domande

Riportiamo un esempio di risposta al *prompt* relativo all'esito della sentenza di primo grado.

- 1 Ti è stato fornito il testo di una sentenza di secondo grado in ambito tributario.
- 2 Puoi fare un riepilogo della decisione presa nel procedimento di primo grado?
- 3 Fornisci le porzioni della sentenza originale da cui hai estratto l'informazione senza apportare modifiche al testo.

Figura 13 – Prompt per il riepilogo dell'esito del primo grado

- 1 Ti è stato fornito il testo di una sentenza di secondo grado in ambito tributario.
- 2 Quali sono i fatti rilevanti nel caso? Includi i dettagli delle questioni di diritto e di fatto contestate nel procedimento di appello.
- 3 Fornisci le porzioni della sentenza originale da cui hai estratto l'informazione senza apportare modifiche al testo.

Figura 14 – Prompt per la specificazione dei fatti rilevanti

- 1 Ti è stato fornito il testo di una sentenza di secondo grado in ambito tributario.
- 2 Qual è la decisione della corte? Riporta le motivazioni su cui si basa la decisione con riferimento a fonti normative e precedenti giurisprudenziali rilevanti. La risposta deve contenere i principi di diritto e i criteri sulla base dei quali la decisione si fonda.
- 3 Fornisci le porzioni della sentenza originale da cui hai estratto l'informazione senza apportare modifiche al testo.

Figura 15 – Prompt per l'indicazione della decisione della Corte

Nel procedimento di primo grado, la Commissione Tributaria Provinciale di Messina aveva accolto i ricorsi proposti da Resistente\_I avverso l'avviso di liquidazione dell'imposta ed irrogazione sanzione con cui l'Ufficio aveva determinato una maggiore imposta di registro di € 15.890,00, oltre interessi e sanzioni a seguito della revoca delle agevolazioni fiscali per l'acquisto della prima casa, e avverso l'avviso di recupero della imposta sostitutiva sulla connessa operazione di mutuo. Il primo Giudice aveva annullato gli atti impugnati, ritenendo che l'agevolazione "prima casa" fosse applicabile al ricorrente, in quanto l'immobile precedentemente acquistato era da considerarsi idoneo rispetto alle esigenze abitative del ricorrente stesso e della di lui famiglia, per la sua consistenza e per le caratteristiche intrinseche.

Porzioni della sentenza originale:

“Con distinte sentenze emesse dalla Commissione Tributaria provinciale di Messina, venivano accolti i ricorsi proposti da Resistente\_I rispettivamente avverso l'avviso di liquidazione dell'imposta ed irrogazione sanzione con cui l'Ufficio aveva determinato una maggiore imposta di registro di € 15.890,00, oltre interessi e sanzioni a seguito della revoca delle agevolazioni fiscali per l'acquisto della prima casa, relativamente all'aggiudicazione dell'immobile sito in Luogol, Indirizza\_1, in catasto al fol 222, part. 87 sub 9; e avverso l'avviso di recupero della imposta sostitutiva sulla connessa operazione di mutuo.”

“Il primo Giudice aveva annullato gli atti impugnati, muovendo dalla considerazione per cui l'agevolazione "prima casa"- disciplinata dall'articolo I della Tariffa — parte prima, nota 11 bis, allegata al D.p.r. n.131/88 - consentiva al ricorrente di usufruire del beneficio fiscale (consistente nell'applicazione dell'imposta di registro con l'aliquota del 2% in luogo del 9%), stante che il ricorrente risultava sì essere proprietario di altra unità immobiliare, acquistato con atto di compravendita registrato a Luogo\_2 P.G. il 19.03.20E13 al n.277, ma l'immobile medesimo era da considerarsi idoneo rispetto alle esigenze abitative del ricorrente stesso e della di lui famiglia, per la sua consistenza e per le caratteristiche intrinseche.”

Nella fase sintetica, le informazioni dettagliate estratte durante la fase analitica sono ricomposte in un sommario coerente e comprensibile. Con il *prompt* si chiede cioè di produrre un sommario unitario che presenti in modo sintetico gli elementi precedentemente individuati nella fase analitica, come si può vedere nella Figura 16.

- 1 Crea una sintesi del testo fornito che contenga:
- 2 - chi ha emesso la sentenza;
- 3 - le parti coinvolte nell'appello e le domande da loro proposte nell'appello;
- 4 - un riepilogo delle decisioni prese nel procedimento di primo grado;
- 5 - i fatti rilevanti nel caso, comprese le questioni di diritto e di fatto contestate nel procedimento di appello;
- 6 - la decisione della corte comprensiva di motivazioni su cui si basa la decisione, riferimenti a fonti normative e precedenti giurisprudenziali rilevanti, principi di diritto e criteri utilizzati.
- 7 Nella sintesi concentrati e dedica più spazio ai principi di diritto.

Figura 16 – Prompt per la sintesi astrattiva combinata

Si riporta un esempio di un sommario astrattivo combinato della sentenza 7683/2022.

La sentenza è stata emessa dalla Corte di Giustizia Tributaria di II grado della Sicilia (ex Commissione Tributaria Regionale della Sicilia) Sezione ID. Le parti coinvolte nell'appello sono l'Agenzia delle Entrate Direzione Provinciale Messina, che propone l'appello, e Resistente\_1, che si oppone all'appello. L'Agenzia delle Entrate chiede di annullare le sentenze di primo grado, mentre Resistente\_1 contesta l'appello a ne chiede il rigetto.

Nel procedimento di primo grado, la Commissione Tributaria Provinciale di Messina aveva accolto i ricorsi proposti da Resistente\_1 avverso gli avvisi di liquidazione emessi dall'Ufficio, che riguardavano la revoca delle agevolazioni fiscali per l'acquisto della prima casa e la conseguente determinazione di una maggiore imposta di registro e il recupero dell'imposta sostitutiva sulla connessa operazione di mutua. La Commissione aveva ritenuto che l'agevolazione "prima casa" fosse applicabile al ricorrente, in quanto (immobile precedentemente acquistato era da considerarsi idoneo rispetto alle esigenze abitative del ricorrente e della sua famiglia.

I fatti rilevanti nel caso riguardano l'impugnazione degli avvisi di liquidazione emessi dall'Agenzia delle Entrate nei confronti del contribuente Resistente1. Le questioni di diritto e di fatto contestate nel procedimento di appello riguardano l'applicabilità delle agevolazioni fiscali per la prima casa all'idoneità dell'immobile pre-acquistato a soddisfare le esigenze abitative del contribuente e della sua famiglia.

La decisione della corte di rigettare il ricorso proposto dall'Ufficio e confermare la sentenza impugnata. La motivazione si basa sulla valutazione dell'idoneità dell'immobile pre-acquistato, considerando il contenuto della perizia giurata a firma del Nominativo\_1, depositata dall'appellato in fase d'appello, che dimostra l'idoneità dell'immobile pre-acquistato a soddisfare le esigenze abitative del ricorrente e della sua famiglia.

## 5.2 Validazione

La validazione dei metodi di sommarizzazione astrattiva tramite LLM si è svolta in due fasi. Nella prima fase sono stati valutati i risultati generati da un ampio numero di modelli (Figura 17) per selezionare quelli migliori e sottoporre a un gruppo di valutatori ridotto solo questi ultimi. In particolare, all'esito della prima fase di validazione, i modelli IT5 avevano ricevuto valutazioni piuttosto negative, mentre GPT-3 e GPT-4 avevano ottenuto ottimi risultati rispetto a tutti i parametri (para. 3.3.2).

	IT5 Small	IT5 Large	GPT3	GPT4	GPT3 a tema	GPT4 a tema
Forma	1.75 (0.97)	1.75 (1.30)	4.12 (0.60)	4.25 (0.66)	3.12 (1.36)	3.62 (1.32)
Completezza	1.62 (1.32)	1.50 (1.00)	4.00 (0.71)	4.00 (0.71)	2.75 (1.20)	3.75 (1.39)
Correttezza	1.62 (0.99)	1.88 (1.54)	4.25 (0.66)	4.25 (0.66)	3.38 (1.32)	3.62 (1.41)
Soddisfazione	1.62 (1.32)	1.50 (1.00)	4.12 (0.33)	4.00 (0.71)	2.62 (1.11)	3.38 (1.32)

Figura 17 – Prima valutazione astrattivi

La seconda fase si è concentrata sulla valutazione dei sommari generati da GPT-4, sia nella versione a testo libero (para. 5.1.1) sia in quella a temi (para. 5.1.2), tenendo IT5 come riferimento per la comparazione. Inoltre, considerati i risultati peggiori rispetto a GPT-4, abbiamo scelto di omettere la valutazione di GPT-3. Dunque, la seconda fase di valutazione si è concentrata su IT5-Small, GPT-4 a testo libero e GPT-4 a tema. La valutazione ha coinvolto un gruppo di circa 80 esperti, composto da giudici, avvocati e altri esperti di diritto. Ogni valutatore ha esaminato 5 decisioni e le relative sintesi prodotte dai 3 modelli. Per ogni modello sono state fornite in media 50 valutazioni.

Come si può vedere nella Figura 18, GPT-4 ha ottenuto un punteggio migliore rispetto a IT5. La sommarizzazione basata sui temi ha avuto punteggi più alti rispetto a quella basata sul testo libero, per quanto riguarda la completezza, la correttezza e la soddisfazione generale. Infatti, il *prompt* astrattivo base opera sul testo libero, quindi senza prendere in considerazione la distinzione tra svolgimento del processo e motivazioni della sentenza. I suoi principali pregi sono la naturalezza e la facilità di fruizione (para. 5.1.1). Il *prompt* per la sommarizzazione a temi si basa invece sull'estrazione dei criteri decisionali (para. 5.1.2), col pregio di produrre

risultati omogenei e basati su informazioni precise e puntuali. Al contempo, secondo gli esperti, la sommarizzazione a testo libero spesso ometteva il riferimento a principi di diritto importanti ai fini della decisione, oppure risultava eccessivamente sintetica. La sommarizzazione a tema, diversamente, era considerata poco conforme agli standard di sintesi cui sono abituati gli operatori giuridici.

	IT5 Small	GPT4	GPT4 a tema
Forma	2.11 (1.10)	3.69 (1.06)	3.69 (1.14)
Completezza	2.15 (1.06)	3.20 (1.25)	3.75 (1.02)
Correttezza	2.51 (1.17)	3.54 (1.15)	3.75 (1.05)
Soddisfazione	2.03 (1.04)	3.30 (1.32)	3.54 (1.12)

Figura 18 – Seconda valutazione astrattivi

La validazione della sommarizzazione astrattiva combinata (para. 5.1.3) ha riguardato un campione di sentenze appartenenti al Dataset n. 3 (para. 3.1) e ha coinvolto un gruppo ristretto di dieci esperti di diritto tributario, cui è stato richiesto di verificare la soddisfazione complessiva dei sommari. È stato espressamente chiesto agli esperti di condurre la loro valutazione tenendo a mente che, come spiegato in precedenza, i sommari non sono riconducibili alla categoria delle “massime”.

La valutazione della soddisfazione complessiva è stata espressa con un punteggio da 1 (riassunto da rigettare) a 3 (riassunto pubblicabile senza modifiche) e ha condotto ad un risultato medio di circa 2.7, con una soddisfazione complessiva superiore a quella riscontrata per i precedenti *output* sottoposti a valutazione.

In particolare, gli esperti hanno sottolineato la completezza e la correttezza dei sommari, evidenziando solo delle imprecisioni lessicali di minore entità.

L'ulteriore e definitivo passaggio di validazione è la sottoposizione dei sommari astrattivi combinati generati sulle sentenze riferibili al Dataset n. 3 ad un gruppo di valutatori selezionati mediante concorso tra studiosi di diritto tributario<sup>40</sup> under 30 e giudici tributari di merito. A questo gruppo è stata proposta la valutazione qualitativa dei sommari astrattivi mediante una piattaforma predisposta *ad hoc* da Sogei. La richiesta era quella di esprimersi sulla base della sola *soddisfazione complessiva* (vedi par. 3.3.2), con la possibilità di inserire modifiche al testo in caso di valutazione inferiore a 3. Inoltre, i commenti con cui i *valutatori* hanno proposto modifiche al testo generato dalla sommarizzazione sono stati valutati puntualmente dal gruppo degli esperti. Anche in questa circostanza è emerso che le correzioni suggerite erano perlopiù relative a mere scelte lessicali e formali, non a errori contenutistici.

## 6 Dalla sommarizzazione alla ricerca semantica

Accanto alla sommarizzazione delle sentenze, PRODIGIT ha esplorato le tecniche per la ricerca semantica. L'obiettivo in questo ambito è di fornire all'utente interessato alla ricerca di sentenze tributarie la possibilità di specificare le *query* in un formato libero, adattabile alle esigenze informative del giurista, ad esempio attraverso l'immissione di stringhe di parole chiave, stralci di sentenze, quesiti giuridici. Sono state sperimentate due diverse tecniche per la ricerca semantica:

- La prima tecnica si basa sulla generazione automatica di parole chiave che descrivono il contenuto della sentenza, poi organizzate in un grafo. Attraverso il grafo si procede alla ricerca delle sentenze rilevanti rispetto al quesito proposto dall'utente (anch'esso descritto mediante parole chiave)
- La seconda tecnica si basa invece sulla creazione di rappresentazioni numeriche (i cosiddetti *embedding*) del significato dei sommari, organizzate in un indice. Le sentenze rilevanti per il quesito proposto dall'utente si ottengono confrontando l'*embedding* del sommario e quello del quesito proposto.

40. In particolare, assegnisti di ricerca in materie tributarie, dottorandi in diritto tributario e laureati che avessero discusso una tesi in diritto tributario (SSD-IUS/12).

Qui ci limitiamo a presentare brevemente la ricerca mediante *embedding*, che illustra la possibilità di usare i sommari non solo per facilitare la comprensione delle sentenze, ma prima di ciò, per consentire la selezione delle sentenze potenzialmente rilevanti.

L'*embedding* di un testo (una parola, una frase o un intero documento) è una struttura matematica (una lista di numeri) che riflette il significato del testo (testi di significato simile hanno *embedding* vicini tra loro). Questa vicinanza viene utilizzata dai sistemi di intelligenza artificiale per cogliere le relazioni semantiche tra parole e utilizzare queste relazioni per compiti come la classificazione del testo, la traduzione automatica, la sintesi vocale, ecc. Per calcolare la similarità tra due testi, mediante il confronto dei loro *embedding*, possono essere utilizzate diverse metriche (in particolare nelle nostre sperimentazioni si è usata soprattutto *la similarità coseno*).

Affidandosi agli *embedding* abbiamo realizzato un sistema di ricerca semantica di sentenze che si basa su quanto segue:

- calcolo degli *embedding* di tutte le sentenze ricercabili (o dei loro sommari, come vedremo);
- calcolo degli *embedding* della *query* dell'utente;
- calcolo della similarità tra l'*embedding* della *query* dell'utente e l'*embedding* delle sentenze nello spazio vettoriale;
- classifica (*ranking*) delle sentenze ordinate secondo il grado di similarità con la *query*;
- restituzione all'utente dell'insieme di sentenze più rilevanti rispetto alla *query* e conseguente determinazione dei documenti più in alto nella classifica.

Questo metodo ha consentito, come vedremo, di ottenere risultati soddisfacenti senza la necessità di un intervento umano (la generazione degli *embedding* è completamente automatica). Non mancano, allo stato dell'arte, i problemi. In primo luogo, il processo di selezione delle sentenze non è trasparente, in quanto la rappresentazione matematica dei testi non è facilmente comprensibile all'essere umano.

In secondo luogo, le tecniche e modelli esistenti per l'ottenimento di una rappresentazione vettoriale a partire dal testo variano molto, dai modelli più semplici (es., vettori di frequenza) a quelli più elaborati (reti neurali e modelli linguistici per l'*embedding*). La selezione del modello migliore per il dominio di riferimento è un passaggio fondamentale. Nel prossimo paragrafo, vedremo come questo problema è stato affrontato nel corso degli esperimenti.

Infine, le complessità intrinseche nel *task* di ricerca delle sentenze sono molteplici. Le sentenze possono essere lunghe decine di pagine, mentre l'utente può limitarsi a ricercare poche parole chiave. Le dimensioni della rappresentazione vettoriale sono limitate ed è quindi necessario assicurarsi che tutte le informazioni essenziali nella sentenza siano conservate in fase di codifica. Solo così il confronto fatto con il vettore generato a partire dalla *query* utente sarà significativo. Come vedremo abbiamo risposto a questa esigenza generando gli *embedding* dei sommari, anziché delle sentenze, creando così una sinergia tra sommarizzazione e ricerca semantica.

## 6.1 Implementazione del prototipo per la ricerca semantica

Sono state verificate sperimentalmente quattro prototipi distinti: (i) uno basato sull'utilizzo di un unico *embedding* per l'intero corpo originale delle sentenze; (ii) uno basato sull'utilizzo di un *embedding* per ogni frase della sentenza originaria; (iii) uno basato sull'uso di un unico *embedding* per il contenuto dell'intero sommario; (iv) l'ultimo che prevedeva quattro distinti *embedding* per i quattro paragrafi del sommario.

La comparazione delle quattro modalità di ricerca è stata effettuata attraverso sessioni interattive che hanno coinvolto gli esperti tributaristi interni al gruppo di lavoro per la sommarizzazione. Agli esperti era richiesto di utilizzare liberamente i quattro prototipi utilizzando diverse modalità di interrogazione: parole chiave, domande libere (es., "Restituisci le sentenze riguardanti l'agevolazione prima casa"), porzioni intere di sentenze. In questo modo è stata presa in considerazione anche la capacità dei sistemi di gestire *input* di lunghezze e

forme completamente diverse fra loro. Agli esperti è stato poi richiesto un giudizio qualitativo sui prototipi che tenesse conto della rilevanza delle prime 10 sentenze ottenute per ogni ricerca e sul loro ordinamento.

I risultati sono stati abbastanza netti: l'utilizzo del quarto metodo di ricerca garantisce risultati comparabili o migliori rispetto agli altri sistemi. Il metodo n. 1 porta a risultati in generale più scadenti, probabilmente a causa della mole di informazioni per singolo *embedding*. Il metodo n. 2 ha mostrato dei risultati in linea di massima deludenti, specialmente in caso di *query* complesse o che coinvolgessero più temi/parole chiave. I metodi n. 3 e n. 4 sono risultati più simili nelle *performance*, ma il n. 4 si è mostrato generalmente più capace di preservare il dettaglio del testo originale e quindi di restituire risultati più attinenti nel caso di ricerche molto specifiche.

Questi risultati non possono essere considerati definitivi, soprattutto vista la ridotta platea di valutatori fino a questo momento coinvolta. Riteniamo, tuttavia, che siano sufficienti per dimostrare le potenzialità dei sommari oltre la semplice fruizione dell'interprete.

I sommari astrattivi in generale, e quelli generati con il metodo combinato in particolare, sono infatti composti solamente dagli elementi di maggiore rilevanza giuridica del testo originale. Questa caratteristica fondamentale li rende particolarmente utili come base per l'uso delle tecniche di *embedding*, in quanto esse saranno applicate solamente agli elementi testuali rilevanti e utili per una ricerca semantica. Inoltre, i sommari combinati possiedono una struttura determinata e una lunghezza limitata, che facilita la generazione di *embedding* che ne colgano adeguatamente il contenuto (più è ampio il testo da cui è tratto l'*embedding*, tanto più è probabile che aspetti importanti del suo significato non abbiano una rappresentazione nell'*embedding* stesso).

La generazione di *embedding* separati per le diverse componenti del sommario ha consentito di ottenere risultati migliori, garantendo una maggiore specializzazione degli *embedding*, e quindi una migliore conservazione del significato di ciascuna componente. Inoltre, questa granularità informativa permette di mantenere insieme degli elementi che potrebbero essere rilevanti nella domanda di ricerca dell'utente, per esempio una particolare catena argomentativa o la presenza di più principi giuridici in fase di decisione.

## 7 Conclusioni

In questo articolo, abbiamo descritto la metodologia e le fasi di sviluppo di un prototipo di IA per la sommarizzazione delle sentenze di merito in materia tributaria. Inoltre, abbiamo presentato come i sommari generati, in particolari quelli relativi alle questioni giuridiche affrontate nella sentenza, possano essere ulteriormente utilizzati come base per la ricerca semantica, attraverso l'uso delle recenti tecniche di NLP basate su processi di *embedding*.

Alla luce di quanto esposto, possiamo qui svolgere qualche considerazione conclusiva.

Innanzitutto, ci sembra che i risultati ottenuti siano accettabili rispetto all'obiettivo che ci siamo proposti, cioè fornire all'operatore del diritto tributario informazioni sufficienti a valutare la rilevanza di una sentenza prima di procedere alla sua lettura integrale. Dunque, allo stadio di evoluzione attuale delle tecnologie, è possibile proporre delle sintesi delle sentenze di merito che siano utili agli operatori del diritto.

In secondo luogo, sembra che gli operatori giuridici preferiscano la sommarizzazione astrattiva delle sentenze alla sommarizzazione estrattiva. Inoltre, essi dimostrano una predilezione per la sommarizzazione astrattiva combinata rispetto alla sommarizzazione a tema.

In terzo luogo, per quanto riguarda la sommarizzazione astrattiva, ci sembra, nel dominio di cui ci siamo occupati, caratterizzato da sentenze tendenzialmente brevi, con contenuti tecnici e considerazioni relative sia al diritto sostanziale, sia procedurale/processuale, che gli LLM a vocazione generalista offrano risultati migliori delle tecnologie di NLP atti a svolgere compiti specifici.

Inoltre, sempre in relazione al dominio di cui si discute, si può affermare che la scelta di chiedere della sommarizzazione astrattiva combinata, che prevede una sorta di frammentazione modulare iniziale del *prompt*, il quale è ricondotto a unità solo in un secondo momento, dia risultati migliori rispetto alla scelta di utilizzare un

unico *prompt*. Ciò sembrerebbe suggerire che la sintesi di testi complessi, come le sentenze, debba avvenire con *prompt* modulari.

Possiamo su questa base formulare alcune considerazioni sugli sviluppi futuri. I metodi proposti sono infatti scalabili e quindi si prestano all'immediata ingegnerizzazione per la realizzazione delle applicazioni operative.

Riteniamo che gli algoritmi e le tecnologie proposte possano essere estese all'intera giurisprudenza tributaria, cosicché ogni sentenza possa essere dotata di sommario che, assieme al testo integrale della sentenza, possa essere messo a disposizione del pubblico. Ciò garantirebbe un migliore accesso alla conoscenza del diritto, da tempo auspicato nel settore fiscale.

Ipotizziamo inoltre che, per estendere la sommarizzazione a domini diversi dall'imposta di registro, saranno utili alcuni raffinamenti delle tecniche da noi utilizzate, in particolare per quanto riguarda i *prompt*. Sebbene non siano state svolte valutazioni approfondite per quanto riguarda l'applicazione del *prompt* utilizzato nel progetto a sentenze relative ad altri settori del diritto sostanziale (es. imposta sui redditi o IVA), sono state svolte alcune analisi e valutazioni sommarie all'interno del gruppo di lavoro. Da queste valutazioni è emerso che i *prompt* dessero buoni risultati anche negli altri settori, ma non così soddisfacenti come nel caso dell'imposta di registro. Ciò potrebbe essere dovuto al fatto che si sono fornite delle lunghezze massime ai sommari, lunghezze adeguate a sentenze relativamente brevi, come nel caso di quelle che componevano il *dataset*. Tale limitazione potrebbe però non essere adeguata ad altre tipologie di sentenze.

Inoltre, si potrebbe immaginare che in ambito tributario, in ragione del principio dell'unitarietà dell'accertamento (para. 1), per poter catturare tutte le informazioni, potrebbe essere opportuno frammentare ulteriormente il *prompt*. In altri termini, si potrebbe immaginare di utilizzare la seguente soluzione: nella prima fase di elaborazione del *prompt* modulare, si dovrebbero inserire ulteriori indicazioni relative al fatto che nella medesima sentenza possono essere trattate questioni relative a imposte diverse e che i criteri decisionali usati dal giudice in relazione alle singole imposte dovrebbero essere estratte separatamente, per essere successivamente assemblate nel sommario unitario.

Quindi, volgendo lo sguardo al di là dei confini tematici e temporali del progetto PRODIGIT, riteniamo che le tecnologie sperimentate nel progetto (e in particolare i modelli linguistici di ampia dimensione) potranno consentire di realizzare importanti perfezionamenti ed estensioni delle funzioni proposte. In particolare, ci sembrano significative le seguenti direzioni per la ricerca e lo sviluppo:

- Miglioramento progressivo delle prestazioni di sommarizzazione (e altre) grazie all'uso di metodologie per il *fine-tuning* automatico. In particolare, nuove tecnologie, da poco disponibili, consentono di far sì che i sistemi apprendano automaticamente ad adeguarsi alle esigenze degli utenti, sulla base delle valutazioni da questi fornite.
- Collegamento con la giurisprudenza della Corte di Cassazione, così fornire l'indicazione non solo di eventuali ricorsi in cassazione, ma anche delle ragioni addotte (o adducibili) a sostegno di accoglimento o rigetto
- Analisi automatiche dell'evoluzione della giurisprudenza tributaria, così da indicare come i criteri di giudizio accolti dai giudici siano mutati nel tempo.