Analisi Preliminare di Conformità e Rischi del Progetto MERL-T/LAIBIT rispetto al Regolamento UE sull'Intelligenza Artificiale (AI Act)

1. Introduzione

Il presente documento analizza il progetto MERL-T (Multi-Expert Retrieval Legal Transformer) e la relativa community LAIBIT (Legal AI Benchmark Italia) alla luce del Regolamento (UE) 2024/1689 ("AI Act"), entrato in vigore il 1° agosto 2024. L'obiettivo è valutare la probabile classificazione di MERL-T secondo l'AI Act, identificare i principali requisiti di conformità applicabili e condurre una valutazione preliminare dei rischi associati.

2. Classificazione di MERL-T secondo l'AI Act

- **Definizione di sistema di IA (Art. 3(1)):** MERL-T rientra chiaramente nella definizione di sistema di IA fornita dall'AI Act, essendo un sistema automatizzato ("machine-based") progettato per operare con autonomia variabile, capace di dedurre ("infer") output (risposte a quesiti giuridici, analisi) da input (query utente, dati), influenzando potenzialmente ambienti virtuali (la conoscenza dell'utente) o fisici (se usato per decisioni con impatto reale). La sua architettura MoA e la capacità di apprendimento (tramite RLCF) ne rafforzano questa classificazione.
- **Valutazione del rischio (Art. 6 e Allegato III):**
  - **Alto rischio - Allegato III, punto 8(a):** la classificazione più probabile per MERL-T è quella di **sistema ad alto rischio**. L'Allegato III, punto 8(a), classifica come tali i sistemi di IA destinati a essere utilizzati da un'autorità giudiziaria o per suo conto per *assistere* nella ricerca e interpretazione dei fatti e del diritto e nell'applicazione del diritto a una fattispecie concreta. Dato che MERL-T mira a fornire risposte contestualmente rilevanti a quesiti giuridici, combinando principi dottrinali/giurisprudenziali (Modulo Principi) e regole normative/fattuali (Modulo Regole), e potenzialmente assistendo nel ragionamento giuridico, il suo utilizzo da parte di professionisti legali (avvocati, magistrati in preparazione di decisioni) o in contesti di ADR rientra pienamente in questa descrizione. Il Considerando 61 conferma questa interpretazione, pur sottolineando il ruolo ausiliario dell'IA.
  - **Altre categorie dell'Allegato III:** A seconda degli specifici moduli o applicazioni future, MERL-T potrebbe tangenzialmente toccare altre aree (es.

Punto 4 - Employment, se usato per valutare performance in studi legali; Punto 6 - Law Enforcement, se integrato con sistemi di analisi predittiva). Tuttavia, la classificazione primaria deriva dal suo scopo di assistenza nel dominio della giustizia.
    - **Deroga ex Art. 6(3):** È **improbabile** che MERL-T possa beneficiare della deroga per sistemi che non pongono un rischio significativo o non influenzano materialmente l'esito decisionale. L'obiettivo stesso di MERL-T è fornire analisi e risposte che *informino e influenzino* il ragionamento e le decisioni legali. Difficilmente potrebbe essere classificato come mero "compito procedurale ristretto" o "preparatorio", specialmente con le sue capacità analitiche e di sintesi. L'eventuale uso di profilazione escluderebbe comunque la deroga.
- **Conclusione sulla classificazione:** MERL-T deve essere considerato, ai fini della conformità, un **sistema di IA ad alto rischio** ai sensi dell'Art. 6(2) e dell'Allegato III, punto 8(a).

3. Requisiti di conformità per MERL-T (sistema ad alto rischio - Art. 8-15)

Essendo un sistema ad alto rischio, MERL-T dovrà soddisfare i rigorosi requisiti del Capo III, Sezione 2 dell'AI Act:

- **Sistema di gestione dei rischi (Art. 9):** È necessario implementare un processo continuo per identificare, valutare e mitigare i rischi per salute, sicurezza e diritti fondamentali. Questo deve considerare errori, bias, usi impropri prevedibili e l'interazione con l'ambiente giuridico complesso. L'approccio RLCF di LAIBIT può contribuire significativamente a questo processo, fornendo un canale per l'identificazione e la mitigazione dei rischi basato sull'esperienza della community.
- **Dati e data governance (Art. 10):** Sfida critica. È necessario garantire che i dati usati per il fine-tuning dei Moduli Principi/Regole (manuali, decisioni Corti Superiori) e per popolare il Vector DB (codici commentati, manuali, massime) siano pertinenti, rappresentativi, privi di errori e completi. Fondamentale è l'analisi e la mitigazione dei **bias** (storici, interpretativi) presenti nei dati legali. L'RLCF può aiutare a identificare e correggere bias tramite feedback esperto. Rimangono aperte le questioni relative ai **diritti d'autore** sui materiali dottrinali/editoriali e alla **privacy** per i dati giurisprudenziali (anche se massimati). Il trattamento di categorie particolari di dati per mitigare i bias è consentito dall'Art. 10(5) a condizioni stringenti.
- **Documentazione tecnica (Art. 11 e Allegato IV):** Sarà richiesto di redigere e mantenere aggiornata una documentazione tecnica estremamente dettagliata,

coprendo architettura, algoritmi, dati, processi di training/testing/validazione, sistema di gestione rischi, misure di sorveglianza umana, performance, etc.
- **Conservazione delle registrazioni (Logging - Art. 12):** MERL-T deve essere progettato per registrare automaticamente eventi rilevanti (es. query, fonti consultate, output generato, interventi umani) per garantire tracciabilità e supportare il monitoraggio post-commercializzazione e le indagini.
- **Trasparenza e informazioni per gli utilizzatori (Art. 13):** Requisito fondamentale nel contesto legale. Le istruzioni per l'uso dovranno spiegare chiaramente le capacità, i **limiti** (incluse accuratezza, robustezza, rischio di "allucinazioni"), i rischi prevedibili, le misure di sorveglianza umana necessarie e come interpretare correttamente gli output. Questo è essenziale per permettere un uso responsabile da parte dei professionisti.
- **Sorveglianza umana (Art. 14):** Il design deve facilitare una sorveglianza umana efficace. L'utente (giurista) deve poter comprendere, monitorare, intervenire, ignorare o bloccare l'output del sistema. Il sistema non deve indurre "automation bias". Questo si allinea con il principio che l'AI assiste ma non sostituisce il giudizio legale.
- **Accuratezza, robustezza, cibersicurezza (Art. 15):** MERL-T deve raggiungere livelli appropriati di accuratezza (da dichiarare), essere robusto contro errori e input malevoli, e garantire la cibersicurezza per proteggere la confidenzialità dei dati legali trattati e l'integrità del sistema stesso (es. contro data poisoning o adversarial attacks).

4. Obblighi per LAIBIT/provider e utilizzatori

- **Provider (LAIBIT/soggetto sviluppatore - Art. 16):** Dovrà implementare un sistema di gestione della qualità (Art. 17), effettuare la valutazione di conformità (Art. 43 - probabilmente interna basata su Annex VI, non richiedendo organismo notificato per Annex III.8a, salvo future modifiche), redigere la dichiarazione di conformità UE (Art. 47), apporre la marcatura CE (Art. 48) e registrare il sistema nel database UE (Art. 49). Dovrà inoltre implementare un monitoraggio post-commercializzazione (Art. 72) e segnalare incidenti gravi (Art. 73).
- **Utilizzatori (Deployers - Art. 26):** Avranno obblighi specifici: usare il sistema secondo le istruzioni, garantire la pertinenza dei dati di input sotto il loro controllo, monitorare il funzionamento, conservare i log (se possibile), implementare la sorveglianza umana e informare i lavoratori se usato in contesto lavorativo.
- **Valutazione d'impatto sui diritti fondamentali (FRIA - Art. 27):** Questo obbligo è **altamente probabile** per molti utilizzatori di MERL-T. Si applica a organismi di

diritto pubblico (es. tribunali, procure, ministeri, università che lo usano per didattica/ricerca con impatto) e a entità private che forniscono servizi pubblici (es. organismi di ADR, forse grandi studi legali in certi contesti). Richiede una valutazione ex-ante dei rischi specifici per i diritti fondamentali nel contesto d'uso e l'adozione di misure di mitigazione. La community LAIBIT potrebbe sviluppare best practice o modelli per supportare gli utilizzatori in questo adempimento.

5. Obblighi di trasparenza specifici (Art. 50)

Indipendentemente dalla classificazione di alto rischio, se MERL-T interagisce direttamente con gli utenti (es. tramite interfaccia chat) o genera contenuti testuali:

- Gli utenti devono essere informati che stanno interagendo con un sistema AI (Art. 50(1)).
- Gli output testuali generati, se pubblicati per informare il pubblico su questioni di interesse pubblico (es. un articolo generato da MERL-T), devono essere etichettati come artificialmente generati, a meno che non vi sia revisione umana con responsabilità editoriale (Art. 50(4)).
- Gli output devono essere marcati in formato leggibile meccanicamente come artificialmente generati (Art. 50(2)), per permetterne il riconoscimento automatico.

6. Considerazioni sui modelli general-purpose (GPAI - Capo V)

Se i Moduli Principi o Regole di MERL-T si basano su modelli fondativi esterni (es. GPT-4, Gemini, Llama 3, modelli open source), il provider di MERL-T diventa un "downstream provider". In questo caso:

- La conformità di MERL-T dipenderà in parte dalla conformità del provider del GPAI a monte (Art. 53), che deve fornire documentazione tecnica (Annex XII), una policy sul copyright e un sommario dei dati di training.
- Se il GPAI sottostante fosse classificato come "a rischio sistemico" (Art. 51), obblighi aggiuntivi (valutazione del modello, mitigazione rischi sistemici, cybersecurity rafforzata - Art. 55) ricadrebbero sul provider del GPAI, influenzando indirettamente MERL-T.

> **7. LAIBIT e RLCF come fattori di mitigazione e conformità**

> **L'approccio collaborativo e validato di LAIBIT, basato su RLCF, si pone come un elemento distintivo che può facilitare la conformità e mitigare alcuni rischi intrinseci:**

- **Bias e qualità dati (Art. 10): Il feedback continuo di esperti diversificati è un meccanismo potente per identificare e correggere bias nei dati e negli output, migliorando la rappresentatività e l'equità.**
- **Gestione rischi (Art. 9): La community agisce come un sistema distribuito di monitoraggio e identificazione dei rischi emergenti durante lo sviluppo e l'uso.**
- **Trasparenza e spiegabilità (Art. 13): L'enfasi sul KG e sulla validazione comunitaria del ragionamento (implicita nell'addestramento del Router/Sintetizzatore) può aumentare la fiducia e potenzialmente la spiegabilità del sistema.**
- **Sorveglianza umana (Art. 14): L'intero processo RLCF è una forma di sorveglianza umana distribuita e qualificata durante la fase di apprendimento e affinamento del sistema.**
- **Supporto all'innovazione (Art. 57): LAIBIT potrebbe gestire o partecipare a sandbox regolatorie per testare MERL-T in modo controllato.**

8. Valutazione preliminare dei rischi residui

Nonostante l'approccio innovativo, permangono rischi significativi:

- **Rischio di conformità:** La complessità e l'onerosità dei requisiti per l'alto rischio (documentazione, QMS, monitoraggio, FRIA per gli utilizzatori) rimangono una sfida, specialmente per una community/progetto potenzialmente basato su risorse limitate o volontarie.
- **Rischio dati:** La dipendenza da fonti esterne (manuali, codici commentati, giurisprudenza) protette da copyright o contenenti dati sensibili rimane un nodo legale ed economico complesso. La qualità e completezza del KG dipendono dalla capacità di estrazione (automatica/manuale) e dalla validazione comunitaria.
- **Rischio di accuratezza/affidabilità:** Anche con RLCF, eliminare completamente "allucinazioni" o errori sottili in output generati da LLM complessi è difficile. Un output legalmente errato può avere conseguenze gravi.
- **Rischio di adozione e governance LAIBIT:** Il successo dipende dalla capacità di costruire e mantenere una community attiva, competente e ben governata. La gestione del feedback RLCF, la risoluzione di disaccordi interpretativi all'interno della community e la definizione della governance del "patrimonio pubblico" (KG, pesi) sono sfide organizzative cruciali.
- **Rischio ecosistema/GPAI:** La dipendenza da GPAI esterni introduce rischi legati alla loro conformità, ai costi, ai cambiamenti di policy dei fornitori. La gestione

dell'ecosistema di terze parti (editori, sviluppatori) richiede accordi chiari.
- **Rischio sicurezza:** La centralità del KG e dei modelli addestrati li rende asset critici da proteggere adeguatamente (cybersecurity, accessi).

9. Conclusione

MERL-T si configura quasi certamente come un sistema di IA ad alto rischio secondo l'AI Act, specificamente nell'ambito dell'amministrazione della giustizia. Ciò comporta l'applicazione di un regime normativo stringente. Il modello LAIBIT, basato sulla validazione comunitaria tramite RLCF e sulla centralità di un Knowledge Graph evolutivo, offre un approccio unico e potenzialmente molto efficace per affrontare alcune delle sfide più complesse poste dall'AI Act, in particolare riguardo alla mitigazione dei bias, alla gestione dei rischi e alla garanzia di affidabilità attraverso la validazione esperta. Questo posiziona il progetto all'avanguardia non solo tecnologicamente, ma anche in termini di governance e allineamento con i principi di trasparenza e affidabilità richiesti dal legislatore europeo e necessari per l'adozione nel settore pubblico. Tuttavia, le sfide legate alla conformità normativa, all'acquisizione dei dati, alla governance della community e alla gestione dell'ecosistema rimangono significative e richiedono un'attenta pianificazione strategica e operativa.