BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Kemajuan teknologi informasi saat ini mendorong banyak organisasi untuk mendigitalisasi dokumen dalam skala besar, termasuk dokumen hukum seperti surat perjanjian. Namun, proses digitalisasi ini masih menghadapi kendala besar karena sistem lama yang digunakan sering kali kaku dan sulit digabungkan dengan teknologi kecerdasan buatan (*Artificial Intelligence* atau AI) terbaru. Hal ini menyebabkan data yang tersedia sulit diolah menjadi informasi yang siap digunakan [1]. Selain itu, pertumbuhan data hukum digital yang sangat pesat atau yang sering disebut sebagai *Legal Big Data* membawa tantangan baru, sebab volume informasi berkembang jauh lebih cepat daripada kemampuan manusia untuk menganalisisnya secara manual [2]. Dokumen tersebut biasanya disimpan dalam format PDF yang tidak terstruktur, sehingga data di dalamnya sulit diproses secara otomatis oleh komputer [3]. Hambatan utama dalam mengelola informasi ini adalah sulitnya mengambil data serta rincian kejadian dari dokumen hukum secara sistematis, yang akhirnya menghambat upaya untuk mengukur efisiensi sistem hukum itu sendiri [1]. Masalah ini diperberat oleh penggunaan bahasa hukum yang rumit dan dokumen yang sangat panjang, sehingga sistem komputer sulit memahami isinya tanpa kehilangan konteks hukum yang penting [4].

Secara spesifik, dokumen perjanjian dagang seperti *Regional Trade Agreements* (RTA) dan *Free Trade Agreements* (FTA) memiliki peran krusial dalam mendorong pertumbuhan ekonomi global dan memfasilitasi akses pasar antar wilayah [5]. Dalam praktiknya, banyak negara mengadopsi pendekatan *boilerplate* atau penggunaan *template* standar dalam menyusun teks legal perjanjian tersebut untuk meminimalkan beban birokrasi [6]. Namun, penggunaan *template* ini tidak bersifat universal terdapat variasi signifikan yang dipengaruhi oleh faktor regional maupun strategi geopolitik, sehingga menciptakan fenomena *legal fragmentation* atau yang dikenal sebagai *spaghetti bowl phenomenon* [6]. Meskipun terdapat kesamaan substansi pada level modular, perbedaan kecil dalam diksi dan struktur antar dokumen menyebabkan proses ekstraksi informasi menjadi sangat kompleks. Ketidakteraturan format visual di balik kesamaan logika hukum inilah yang menciptakan kebutuhan akan pendekatan ekstraksi berbasis *Large Language Model*

yang mampu memahami semantik teks secara fleksibel tanpa bergantung pada aturan tata letak yang kaku.

Meskipun *Large Language Model* (LLM) generasi terbaru memiliki *context window* yang sangat besar hingga mencapai satu juta token, efektivitasnya dalam pemrosesan dokumen hukum yang panjang tetap menghadapi batasan kognitif yang signifikan. Penelitian terbaru menunjukkan adanya degradasi akurasi ekstraksi sebesar 8 hingga 10 poin persentase seiring dengan meningkatnya panjang konteks dokumen [7]. Fenomena ini diperparah dengan adanya *positional bias*, di mana informasi yang terletak di bagian tengah dokumen cenderung lebih sulit diidentifikasi dibandingkan informasi di bagian awal atau akhir [7]. Selain masalah jangkauan konteks, LLM juga menunjukkan keterbatasan dalam menjaga konsistensi urutan logis (*sequencing*) dan struktur data saat dihadapkan pada tugas yang memerlukan penalaran multi langkah (*multi step reasoning*) [7, 8]. Ketidakmampuan model untuk mempertahankan integritas skema data secara mandiri dalam dokumen berskala besar menjustifikasi kebutuhan akan mekanisme orkestrasi seperti *Recursive Character Splitting*. Dengan membagi dokumen menjadi segmen yang lebih kecil namun tetap mempertahankan hubungan semantik, tantangan *contextual loss* dapat diminimalisir guna menjamin hasil ekstraksi yang akurat dan *machine processable*.

Upaya untuk mengatasi limitasi pemrosesan dokumen legal telah mendorong pengembangan sistem orkestrasi yang mengintegrasikan LLM dengan infrastruktur manajemen data terstruktur. Salah satu implementasi yang relevan adalah sistem *Contrato360*, yang berhasil menggabungkan teknik *Retrieval Augmented Generation* (RAG), *Text to SQL*, *Prompt Engineering*, dan arsitektur *multi agent* untuk menjawab pertanyaan pengguna berdasarkan kombinasi data dari dokumen kontrak PDF dan sistem manajemen kontrak yang sudah terstruktur [9]. Meskipun pendekatan ini terbukti efektif untuk kontrak bisnis berdurasi pendek, terdapat tiga keterbatasan fundamental yang belum terpecahkan. Pertama, Contrato360 beroperasi dalam paradigma *Question Answering* (QA) dan tidak melakukan transformasi dokumen mentah menjadi basis data relasional dari nol; sistem ini bergantung pada keberadaan basis data yang sudah terstruktur sebelumnya sebagai salah satu sumber informasi. Kedua, dokumen kontrak yang diproses Contrato360 memiliki struktur datar (*flat structure*) dan berdurasi pendek, sehingga tidak menghadapi tantangan degradasi akurasi akibat *contextual loss* dan *positional bias* yang krusial pada dokumen hukum berdimensi panjang seperti RTA. Ketiga, Contrato360 tidak menerapkan mekanisme *chunking* dan *output parsing*

ketat untuk menjamin konsistensi skema data relasional yang diekstraksi dari dokumen tidak terstruktur.

Penelitian terbaru oleh [10] membuktikan bahwa penggunaan model seperti GPT-4o dalam skema ekstraksi hierarkis mampu menghasilkan akurasi QA hingga 97.5%, namun tantangan inkonsistensi skema (*schema inconsistency*) tetap muncul ketika *output* AI menyimpang dari format yang telah ditentukan. Di sisi lain, penelitian pada domain kecelakaan lalu lintas menunjukkan bahwa LLM masih menghadapi kendala serius dalam menjaga konsistensi urutan peristiwa (*event sequencing*) dan integritas struktur data pada dokumen naratif yang panjang, bahkan setelah optimasi *prompt* [8]. Kedua temuan ini mengonfirmasi bahwa diperlukan sebuah *hybrid pipeline* yang tidak hanya mengandalkan *prompting*, tetapi juga mengintegrasikan kemampuan pemahaman semantik LLM dengan mekanisme orkestrasi yang ketat melalui *framework* seperti LangChain. Pendekatan ini memungkinkan transformasi dokumen RTA yang memiliki struktur hierarkis mendalam dan rentan terhadap *contextual loss* menjadi basis data relasional yang konsisten dan andal, sekaligus menjembatani kesenjangan antara paradigma QA yang berfokus pada *downstream retrieval* (seperti Contrato360) dan kebutuhan *upstream digitization* dari dokumen legal mentah menjadi data *machine processable*.

Upaya terbaru dalam mentransformasikan dokumen hukum perdagangan telah mulai mengeksplorasi penggunaan LLM untuk pembentukan graf pengetahuan (*Knowledge Graph*) secara otomatis dari teks RTA [11]. Penelitian tersebut menunjukkan bahwa teknik *prompting* yang tepat mampu mengekstraksi hubungan antar entitas hukum dengan tingkat keberhasilan yang menjanjikan, namun masih menghadapi tantangan dalam hal konsistensi format dan pemetaan skema yang kaku (*schema enforcement*) [11]. Selain itu, tantangan teknis dalam menerjemahkan narasi hukum menjadi *Structured Query Language* (SQL) atau skema tabel relasional seringkali terkendala oleh ambiguitas terminologi dan kompleksitas hubungan *Parent-Child* antar pasal [12]. Tanpa mekanisme pemetaan yang presisi, data yang diekstraksi berisiko kehilangan integritas hierarkisnya saat disimpan ke dalam sistem basis data relasional.

Berdasarkan celah penelitian tersebut, penelitian ini mengusulkan implementasi sistem transformasi dokumen legal *end to end* yang spesifik menangani dokumen *Trade Agreement* menggunakan *Large Language Model* dan LangChain. Kebaruan penelitian ini terletak pada integrasi teknik *Recursive Character Splitting* untuk menjaga konteks dokumen panjang dan penggunaan

output parsing yang ketat untuk menjamin integritas data saat ditransformasikan ke dalam basis data relasional. Berbeda dengan pendekatan *Question Answering* umum [9, 10], sistem ini berfokus pada digitalisasi struktur hierarkis perjanjian dagang agar menjadi data yang *machine processable*. Lebih jauh, penelitian ini merupakan langkah awal dari visi jangka panjang untuk membangun sistem pencarian semantik berbasis *Retrieval Augmented Generation* (RAG) pada korpus dokumen *Trade Agreement*, di mana kualitas fondasi data relasional yang dihasilkan pada tahap ini akan menentukan efektivitas seluruh *pipeline* pencarian di tahap selanjutnya.

Untuk mencapai tujuan tersebut, tiga hipotesis penelitian dirumuskan guna menguji hubungan antara variabel teknis dan metrik evaluasi. Pertama, variasi ukuran fragmen (*chunk size*) pada teknik *Recursive Character Splitting* berpengaruh signifikan terhadap *Mean Absolute Error* (MAE) ekstraksi elemen hierarkis dokumen *Trade Agreement*; semakin kecil ukuran *chunk*, MAE akan semakin rendah karena setiap segmen teks lebih mudah dipahami secara utuh oleh LLM tanpa mengalami fenomena *lost in the middle*. Kedua, variasi persentase tumpang tindih (*chunk overlap*) pada teknik *Recursive Character Splitting* berpengaruh signifikan terhadap akurasi ekstraksi; semakin besar persentase *overlap*, akurasi akan semakin tinggi karena informasi yang terpotong di batas antar *chunk* dapat direkonstruksi. Ketiga, penggunaan mekanisme *sequential chaining* yang menyertakan konteks fragmen sebelumnya (*previous context*) mampu menurunkan MAE secara signifikan dibandingkan pemrosesan *chunk* secara independen, khususnya pada dokumen dengan struktur hierarkis mendalam seperti *Chapter* $\rightarrow$ *Section* $\rightarrow$ *Article*. Pengujian terhadap ketiga hipotesis ini dilakukan melalui studi ablasi dengan memvariasikan parameter *chunk size* dan *chunk overlap*, serta membandingkan performa antara pemrosesan dengan dan tanpa *sequential chaining* menggunakan MAE sebagai metrik utama dan akurasi sebagai metrik pendukung.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan sebelumnya, rumusan masalah dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana mengimplementasikan *framework* LangChain pada Large Language Model untuk mentransformasikan dokumen *Trade Agreement* yang tidak terstruktur menjadi basis data relasional yang *machine processable*?

2. Bagaimana performa penggunaan teknik *Recursive Character Splitting* dan *sequential chaining* dalam menjaga integritas struktur hierarki (perjanjian, pasal, dan ayat) pada dokumen hukum berdimensi panjang?

1.3 Batasan Permasalahan

Dalam penelitian ini, terdapat beberapa batasan permasalahan yang perlu diperhatikan untuk menjaga fokus dan ruang lingkup penelitian. Batasan-batasan tersebut antara lain:

1. Dokumen sumber yang diproses terbatas pada dokumen legal berformat *Digital PDF* (*text based*), sehingga penelitian ini tidak mencakup pemrosesan dokumen hasil pemindaian (*scanned document*) yang memerlukan *Optical Character Recognition* (OCR) tingkat lanjut.

2. Dataset yang digunakan terdiri dari 10 dokumen *Regional Trade Agreements* (RTA) terpilih dalam bahasa Inggris, dengan fokus pada ekstraksi struktur logis teks dan bukan pada analisis interpretasi hukum spesifik suatu yurisdiksi.

3. Mekanisme pemrosesan dokumen panjang dilakukan secara sekuensial menggunakan metode *chunking* dan *chaining* pada *framework* LangChain tanpa menggunakan *Vector Database*. Hal ini dilakukan untuk menjamin integritas urutan informasi (*contextual sequence*) yang krusial bagi pembentukan basis data relasional.

4. Penelitian ini berfokus pada dokumen yang mengikuti kaidah penulisan perjanjian internasional standar yang memiliki elemen hierarki (Perjanjian, Pasal, dan Ayat), dan tidak menangani dokumen dengan kerusakan struktur teks atau artefak digital yang ekstrem.

5. Skema basis data relasional yang dihasilkan terbatas pada entitas utama dokumen perjanjian dagang dan tidak mencakup seluruh meta data teknis yang tidak relevan dengan struktur hukum inti.

6. Proses ekstraksi informasi hanya berfokus pada konten tekstual naratif (*pure text*), sehingga elemen non-tekstual seperti grafik, diagram, serta data di dalam tabel yang kompleks tidak termasuk dalam cakupan transformasi ke basis data relasional.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

1. Merancang dan membangun arsitektur sistem *end to end* yang mampu mentransformasikan dokumen *Trade Agreement* dari format PDF tidak terstruktur menjadi basis data relasional yang *machine processable* secara otomatis menggunakan *framework* LangChain dan *Large Language Model*.
2. Mengevaluasi dan mengimplementasikan teknik *Recursive Character Splitting* serta *sequential chaining* untuk menjaga akurasi ekstraksi dan integritas struktur hierarki (hubungan perjanjian, pasal, dan ayat) pada dokumen hukum berdimensi panjang.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan kontribusi signifikan, baik dari aspek akademis maupun praktis, khususnya dalam ranah informatika hukum (*legal informatics*). Berikut adalah beberapa manfaat yang dapat diperoleh dari penelitian ini:

1. Menjadi referensi ilmiah mengenai efektivitas penggunaan *Large Language Model* dan *framework* LangChain dalam menangani tantangan ekstraksi informasi semantik dari dokumen dengan konteks panjang dan struktur hierarki yang kompleks.
2. Menjadi acuan dalam penerapan metode *advanced layout parsing* menggunakan Docling untuk mengenali struktur visual dan hierarki pada dokumen legal internasional seperti *Regional Trade Agreements* (RTA).
3. Memberikan solusi bagi organisasi internasional, instansi pemerintah, atau praktisi hukum untuk mentransformasikan dokumen legal tidak terstruktur menjadi basis data relasional (PostgreSQL) secara otomatis dan akurat.
4. Meningkatkan efisiensi dalam pengelolaan dokumen hukum berskala besar dengan meminimalisir risiko kesalahan input data manual (*human error*) serta mempercepat proses pencarian informasi melalui mekanisme *query* basis data yang terorganisir.

5. Menjadi rujukan bagi pengembangan sistem manajemen pengetahuan hukum yang lebih cerdas, mendukung proses pengambilan keputusan berbasis data, dan mempercepat digitalisasi arsip hukum internasional.

1.6 Sistematika Penulisan

Berisikan uraian singkat mengenai struktur isi penulisan laporan penelitian, dimulai dari Pendahuluan hingga Simpulan dan Saran.

Sistematika penulisan laporan adalah sebagai berikut:

1. Bab 1: PENDAHULUAN
   Bab ini berisi latar belakang masalah, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan laporan.

2. Bab 2: LANDASAN TEORI
   Bab ini berisi tinjauan pustaka dan landasan teori yang relevan, meliputi penelitian pendahulu, konsep *Large Language Model* (LLM), teknik konversi dokumen menggunakan *Docling*, orkestrasi sistem dengan *Framework LangChain*, tinjauan mengenai dokumen *Regional Trade Agreement*, serta evaluasi sistem menggunakan *mean absolute error*.

3. Bab 3: METODOLOGI PENELITIAN
   Bab ini menjelaskan metodologi penelitian yang digunakan, analisis kebutuhan sistem, serta perancangan sistem yang dikembangkan, mencakup *flowchart* sistem, perancangan skema basis data, dan struktur tabel untuk penyimpanan data hukum terstruktur.

4. Bab 4: HASIL DAN DISKUSI
   Bab ini membahas implementasi sistem yang dikembangkan, pengujian sistem, serta hasil pengujian dan evaluasi yang dilakukan untuk menjawab rumusan masalah penelitian.

5. Bab 5: SIMPULAN DAN SARAN
   Bab ini berisi kesimpulan dan saran yang diperoleh dari penelitian serta saran untuk pengembangan penelitian selanjutnya.