BAB I

PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi *Artificial Intelligence* (AI) dalam beberapa tahun terakhir telah menunjukkan kemajuan yang signifikan, khususnya pada model-model *multimodal* yang mampu mengintegrasikan pemahaman teks dan gambar [1], [2]. Setelah hadirnya *Generative AI* berbasis *text-to-image* (T2I) seperti *Stable Diffusion* dan *Gemini*, berkembang pula *Vision-Language Model*, yaitu sistem AI yang mampu menganalisis konten gambar, mengidentifikasi objek, serta menjawab pertanyaan berdasarkan konteks visual [3], [4]. Teknologi ini memiliki potensi besar dalam bidang pendidikan, terutama dalam menyediakan pengalaman belajar interaktif melalui kombinasi visual dan penjelasan berbasis bahasa natural [5]. Berdasarkan teori *Multimedia Learning*, penyajian teks dan visual secara simultan dapat meningkatkan pemrosesan kognitif serta memperkuat retensi informasi pada siswa sekolah dasar [13].

Dalam konteks lokal, Desa Wisata Tigaraksa memiliki potensi sebagai pusat edukasi berbasis pertanian dengan lingkungan belajar yang dekat dengan kehidupan sehari-hari anak-anak. Namun demikian, pengembangan materi ajar visual di wilayah ini masih menghadapi berbagai tantangan, khususnya keterbatasan sumber daya desain dan fotografi. Penelitian sebelumnya telah menawarkan solusi awal melalui pemanfaatan teknologi T2I untuk menghasilkan gambar edukatif berbasis AI yang merepresentasikan konteks lokal Desa Tigaraksa. Melalui perbandingan metode *prompt engineering* (*descriptive*, *instruction-based*, dan *compositional*) pada dua model T2I, penelitian tersebut membantu guru dalam menghasilkan visual pembelajaran yang relevan.

Meskipun demikian, pemanfaatan gambar dalam pembelajaran tidak berhenti pada tahap pembuatan visual semata. Dalam praktik pembelajaran, gambar perlu dipahami, dijelaskan, serta dijadikan dasar interaksi antara guru dan siswa. Guru

membutuhkan dukungan sistem untuk memberikan penjelasan berbasis visual, sementara siswa sekolah dasar memerlukan pendamping belajar yang mampu menjawab pertanyaan sederhana terkait isi gambar [11]. Hingga saat ini, belum tersedia sistem yang mampu menghubungkan gambar-gambar lokal hasil T2I dengan penjelasan berbasis AI yang relevan, aman, dan sesuai dengan tingkat pemahaman siswa. Dengan kata lain, penelitian sebelumnya belum mengakomodasi aspek pemahaman *multimodal* dan penjelasan visual secara interaktif [10].

Kesenjangan tersebut menunjukkan perlunya penelitian lanjutan yang tidak hanya berfokus pada generasi gambar, tetapi juga pada pengembangan sistem yang mampu memahami dan menalar konten visual. Terdapat tiga *research gap* utama yang perlu dijawab [6],[9]. Pertama, belum tersedia sistem yang secara sistematis mengekstraksi *visual metadata* dari gambar edukatif dan memanfaatkannya sebagai dasar *reasoning* [12]. Kedua, belum dikembangkan *multimodal chatbot* yang dirancang khusus untuk konteks pembelajaran berbasis desa yang dapat mendukung guru maupun siswa melalui percakapan berbasis visual [10]. Ketiga, belum terdapat integrasi antara *vision model* dengan pendekatan *Retrieval-Augmented Generation* (RAG) dan *Small Language Model* (SLM) untuk memastikan jawaban chatbot tetap akurat, tidak mengalami *hallucination*, dan sesuai dengan konteks gambar [18], [19].

Untuk menjawab kesenjangan tersebut, penelitian ini mengusulkan pengembangan *Multimodal Chatbot* berbasis *Vision AI* dan RAG dengan mengintegrasikan model *BLIP* (Salesforce) untuk menghasilkan *caption* dan deskripsi visual [3], *Sentence Transformer* untuk membentuk *embedding* dan melakukan pencarian vektor [2], serta *Optical Character Recognition* (OCR) untuk mengekstraksi teks yang muncul pada gambar. Seluruh informasi visual tersebut disimpan secara terstruktur dalam *Supabase* sebagai basis pengetahuan. Pada tahap *reasoning*, chatbot menggunakan *Small Language Model* (*GPT OSS*) untuk menghasilkan respons yang terkontrol [34]. Pendekatan RAG memastikan bahwa setiap jawaban yang dihasilkan selalu merujuk pada metadata gambar yang

relevan, seperti *caption*, objek visual, teks OCR, dan hasil *embedding*, sehingga respons bersifat faktual, aman, dan sesuai dengan konteks pembelajaran [35]. Pipeline ini memungkinkan chatbot memanfaatkan kombinasi metadata visual sebagai konteks percakapan dalam menjawab pertanyaan guru maupun siswa secara konsisten dan akurat.

Penelitian ini memberikan kontribusi teoretis dengan memperkaya literatur mengenai integrasi *Vision AI*, RAG, dan SLM dalam konteks pendidikan dasar [1], [6], [13], serta memperkenalkan *multimodal pipeline* yang efisien berbasis model *open-source* [4]. Secara praktis, penelitian ini menghasilkan prototipe chatbot edukatif yang membantu guru dan siswa di Desa Wisata Tigaraksa dalam memahami konten visual lokal melalui percakapan natural, sehingga memudahkan akses terhadap materi visual yang lebih informatif, terarah, dan sesuai dengan kebutuhan pembelajaran [5], [11].

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana kinerja model vision open-source (BLIP dan OCR) dalam mengekstraksi metadata visual dari gambar edukatif lokal Desa Wisata Tigaraksa?
2. Bagaimana proses embedding menggunakan *Sentence Transformer* dapat digunakan untuk melakukan pencarian vektor yang akurat dalam sistem RAG?
3. Bagaimana *Small Language Model (GPT OSS)* mampu menghasilkan jawaban edukatif yang relevan, aman, dan sesuai konteks gambar ketika dipadukan dengan RAG?

1.3. Tujuan Penelitian

Penelitian ini bertujuan untuk:

1. Mengembangkan sistem chatbot multimodal berbasis *Vision AI, OCR, Sentence Transformer* dan *Small Language Model* yang mampu memahami gambar.
2. Mengimplementasikan dan menguji *pipeline* RAG untuk memastikan setiap jawaban chatbot berdasarkan pada metadata visual yang benar dan tidak halusinasi.
3. Mengevaluasi performa model vision (BLIP + OCR) dalam menghasilkan caption, objek dan teks yang relevan untuk pembelajaran anak-anak desa.
4. Menilai kualitas respons chatbot melalui uji pengguna (guru dan siswa SD Tigaraksa) berdasarkan aspek relevansi, kejelasan dan kesesuaian edukatif.
5. Menghasilkan prototipe chatbot edukatif yang dapat diterapkan dalam program pembelajaran berbasis gambar di Desa Wisata Tigaraksa.

1.4. Urgensi Penelitian

Penelitian ini memiliki urgensi karena kebutuhan media pembelajaran interaktif di Desa Wisata Tigaraksa belum sepenuhnya terpenuhi, terutama dalam hal pemahaman gambar edukatif yang sesuai dengan konteks lokal. Keterbatasan tenaga desain dan materi visual membuat guru kesulitan memberi penjelasan yang terstruktur kepada siswa, sementara anak-anak membutuhkan pendamping belajar yang mampu menjawab pertanyaan berbasis visual secara sederhana dan aman. Di sisi lain, perkembangan teknologi AI multimodal termasuk *Vision AI, OCR, Sentence Transformer* dan *Small Language Model* membuka peluang besar untuk menghadirkan sistem pembelajaran yang lebih cerdas dan mudah diakses. Oleh karena itu, penelitian ini penting untuk menjembatani kesenjangan antara ketersediaan teknologi dan kebutuhan pembelajaran di lingkungan desa, sekaligus mendorong pemanfaatan AI yang tepat, aman dan berkelanjutan dalam konteks pendidikan dasar.

1.5. Luaran Penelitian

Luaran penelitian ini mencakup pengembangan sebuah prototipe chatbot multimodal berbasis *Vision AI* dan RAG yang mampu memahami gambar edukatif dan memberikan jawaban yang relevan bagi guru maupun siswa.

1.6. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat baik secara teoritis maupun praktis yang signifikan bagi pengembangan ilmu pengetahuan. Berikut adalah manfaat penelitian secara teoritis dan praktis

1. Manfaat Teoretis
- Menambah literatur mengenai integrasi *Vision AI, OCR, Sentence Transformer,* dan *RAG* untuk konteks pendidikan dasar.
- Menghasilkan model konseptual pipeline multimodal yang dapat direplikasi untuk studi sejenis.
2. Manfaat Praktis
- Menyediakan alat bantu pembelajaran berbasis pembelajaran bagi guru dan siswa Desa Wisata Tigaraksa.
- Mempermudah siswa dalam memahami gambar melalui penjelasan yang natural dan relevan.
- Mendukung program pembelajaran desa berbasis teknologi dan memperkuat akses pendidikan digital.
- Memberikan dasar teknologi untuk pengembangan platform pembelajaran visual berbasis komunitas.