

OMem: Dynamic Hybrid Retrieval for Long-Term Memory in LLM-Based Agentic Systems

Ziad El Sayed

MSc Machine Learning dissertation, University College London, September 2025

Supervisors: Dr Omer Gunes (University of Oxford), Prof Philip Treleaven (UCL)

Introduction

Despite the expansion of context windows in frontier models, the inability to recall earlier sessions, facts, and goals continues to undermine their usefulness in long-term or agentic tasks [1].

Dense retrievers may miss exact matches [2], while sparse retrievers may struggle with paraphrased inputs [3].

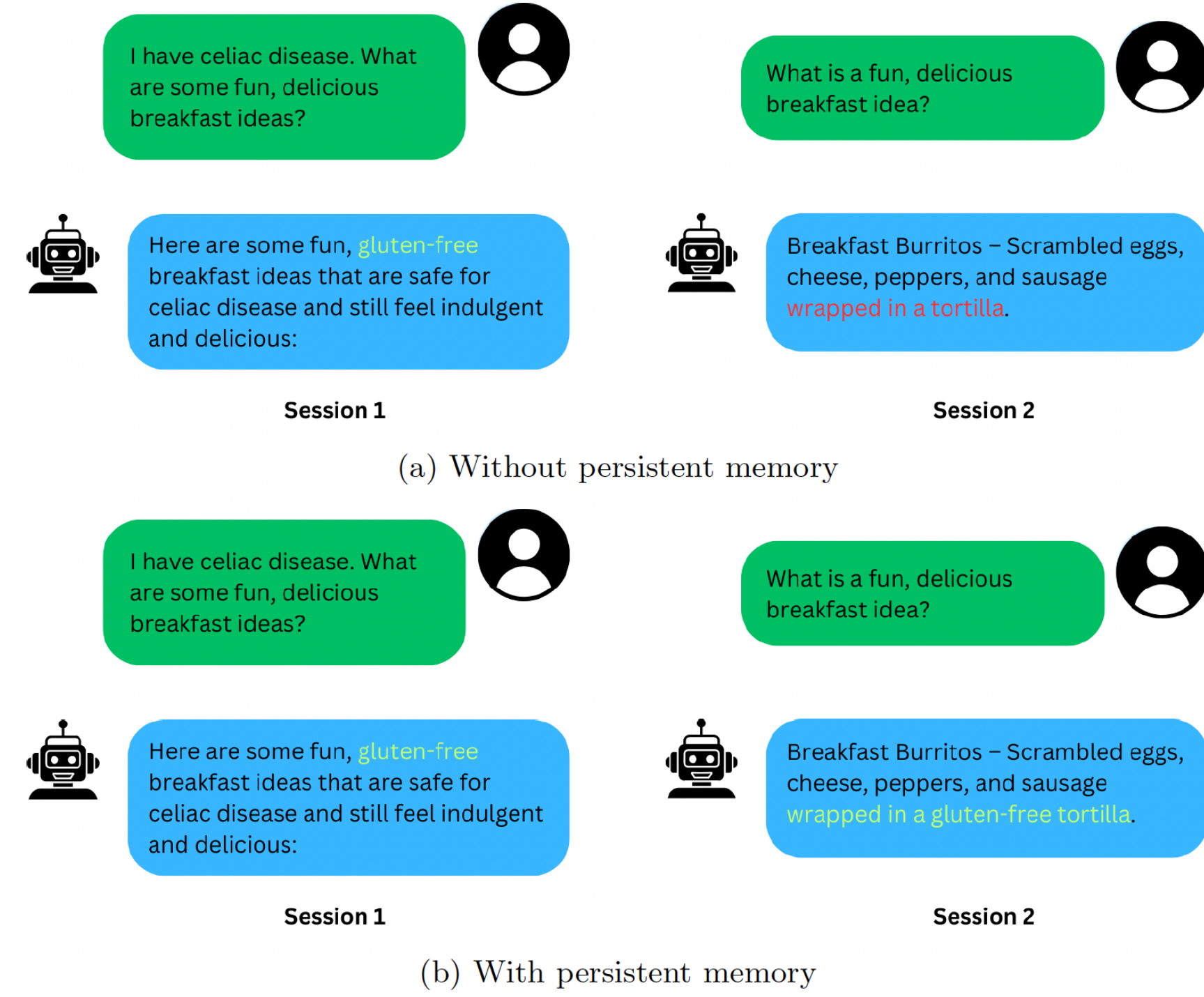


Figure 1: Comparison of LLM behaviour with and without memory continuity.

Approach

This thesis investigates dynamic memory systems for agentic LLMs. It introduces OMem, a modular, hybrid memory architecture that combines dense and sparse retrieval methods, integrates reciprocal rank fusion (RRF) [4], and uses cross-encoder reranking to filter retrieved memories before reasoning.

This thesis makes several contributions to the field of memory-augmented large language models:

1. A modular hybrid retrieval architecture (OMem) that supports configurable combinations of dense retrieval, sparse lexical matching, reciprocal rank fusion, and cross-encoder reranking.
2. Fine-grained empirical quantification of the individual and combined impact of retrieval methods, supported by targeted ablation studies.
3. A reproducible evaluation suite using long-term memory benchmarks (LoCoMo [5] and LongMemEval [6]).
4. Empirical validation demonstrating state-of-the-art performance on long-term QA tasks.

Method

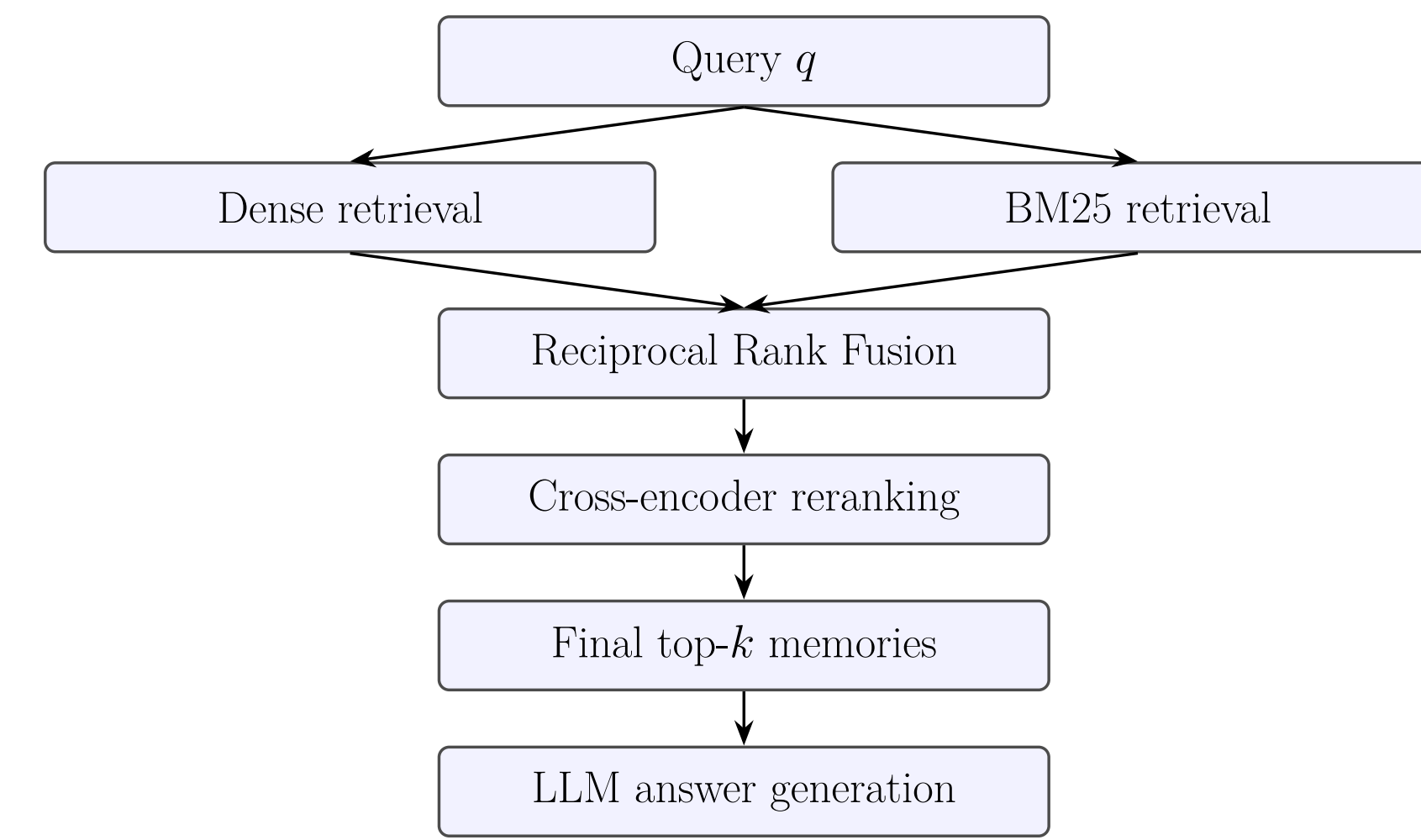


Figure 2: OMem retrieval and reranking pipeline. A query is retrieved against dense and sparse indexes, fused with RRF, and reranked with a cross-encoder to produce final top- k memories for the LLM.

1. Semantic similarity (dense retrieval)

Dense embeddings represent both conversational utterances and user queries as continuous vectors, in which geometric proximity corresponds to semantic similarity:

$$R_{\text{dense}}(q) = \arg \text{top-}k \text{ sim}(q, u_i)$$

2. Lexical similarity (BM25)

BM25 preserves exact lexical matches such as names, numbers, or dates that dense embeddings may blur:

$$R_{\text{sparse}}(q) = \arg \text{top-}k \text{ BM25}(q, u_i)$$

3. Reciprocal rank fusion

RRF merges retrieval lists into a single consensus ranking, from the positions $r_{\text{dense}}(d)$ and $r_{\text{sparse}}(d)$ of document d in the two lists:

$$\text{RRF}(d) = (1 - \lambda) \frac{1}{k + r_{\text{dense}}(d)} + \lambda \frac{1}{k + r_{\text{sparse}}(d)}$$

where $k = 60$ dampens the effect of deep ranks and $\lambda \in [0, 1]$ controls the contribution of each retriever.

4. Cross-encoder reranking

A cross-encoder jointly encodes the query-candidate pair and outputs a direct relevance score; OMem applies it only to the top 100 candidates and keeps the top 50:

$$s_{\text{CE}}(q, u) = f_{\phi}([q; u])$$

where f_{ϕ} is a transformer with parameters ϕ .

Effect of λ_{sparse}

On LoCoMo, $\lambda_{\text{sparse}} \in [0.0, 0.3]$ represents the optimal operating range across different retrieval budgets. LongMemEval benefits more clearly from incorporating lexical information, especially at moderate fusion weights.

Results

Table 1: Overall Results on LoCoMo (comparison with MemO [7]).

Method	BLEU	F1	LLM
MemO (Prev. SOTA)	0.3696	0.4650	0.6792
OMem (GPT-4o-mini)	0.4122	0.5220	0.7552
OMem (GPT-4o)	0.4422	0.5690	0.7916

MemO's [7] answer generation and grading is performed on GPT 4o-mini, as done by the original authors.

The overall LLM-as-a-Judge score of 0.79 achieved by OMem on GPT-4o substantially surpasses previously reported results and establishes a new state of the art.

Table 2: Overall Results on LongMemEval (comparison with Zep [8]).

Method	Accuracy	Latency (s)
Zep (GPT-4o-mini)	0.638	3.20
OMem (GPT-4o-mini)	0.670	1.84
Zep (GPT-4o)	0.712	2.58
OMem (GPT-4o)	0.766	2.54

We state the results as reported by the Zep team in their report [8].

OMem surpasses Zep across both model scales. OMem achieves near-perfect accuracy on single-session assistant (98%) and single-session user (93%), substantially outperforming Zep, which drops to 75–80% on assistant questions.

Ablations

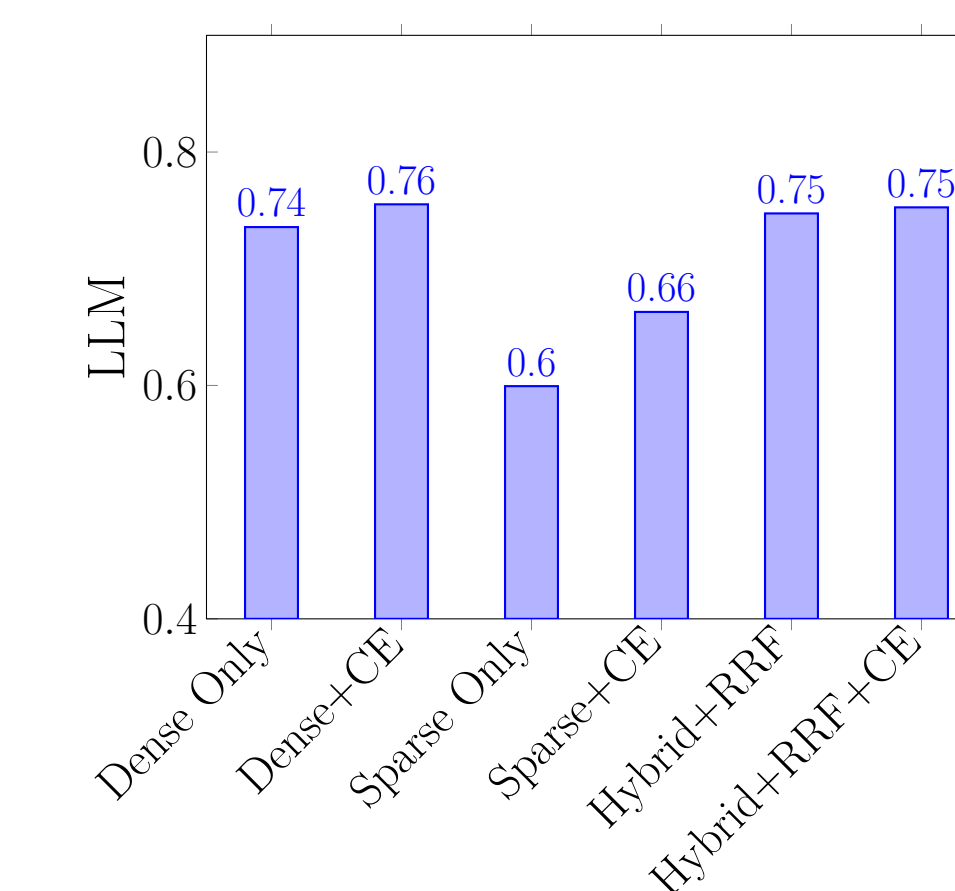


Figure 3: Overall LLM-as-a-Judge scores across ablation configurations on LoCoMo.

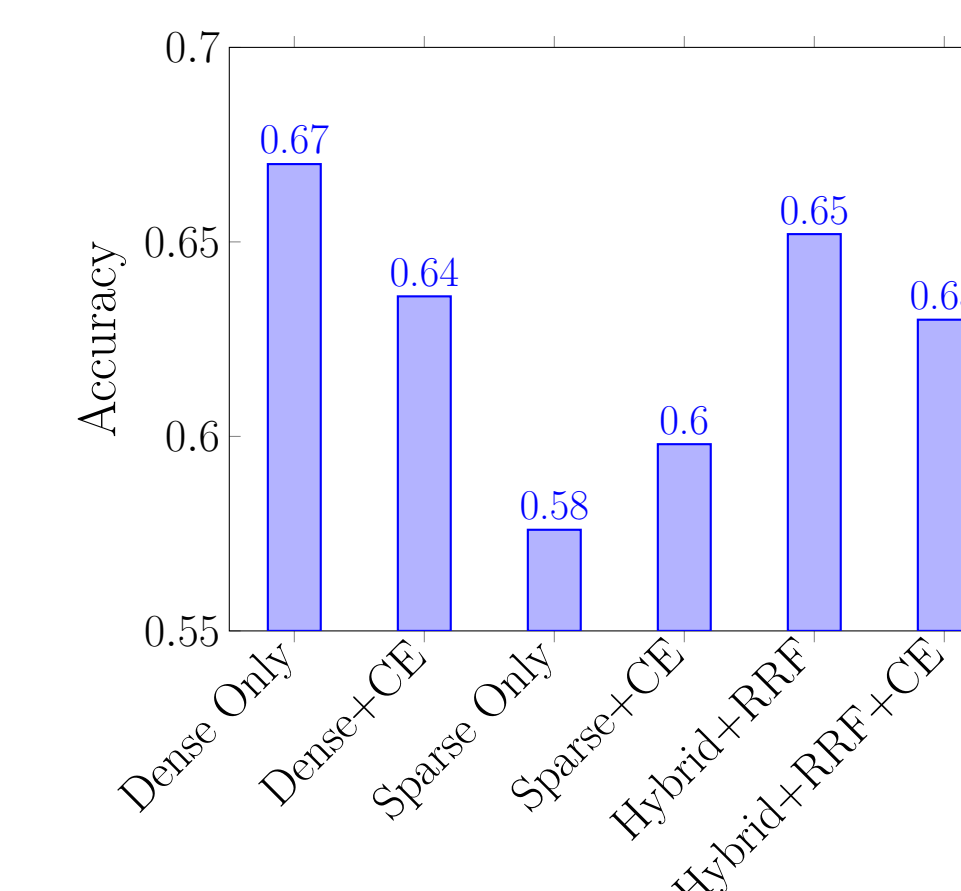


Figure 4: Overall accuracy across ablation configurations on LongMemEval.

Limitations

- Zep retains an advantage on preference questions, achieving 53–57% compared to OMem's 17–50%. This highlights a remaining challenge for OMem in capturing implicit, subjective user preferences.

Conclusions

- Direct retrieval-based memory significantly outperforms summarisation-based systems.
- Dense retrieval consistently outperformed sparse and hybrid approaches across benchmarks.
- Cross-encoder reranking boosts performance in some settings.
- Hybrid retrieval is not universally beneficial and is sensitive to tuning.
- Modular memory architectures like OMem allow system designers to make informed, component-level trade-offs between latency and retrieval fidelity.

References

- [1] Huang et al. (2025). Language models do not have human-like working memory. arXiv:2505.10571.
- [2] Chen et al. (2022). Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? Findings of EMNLP 2022, 250–262.
- [3] Shehata, Arabzadeh and Clarke (2022). Early stage sparse retrieval with entity linking. CIKM 2022, 4464–4469.
- [4] Cormack, Clarke and Buettcher (2009). Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. SIGIR 2009, 758–759.
- [5] Maharana et al. (2024). Evaluating very long-term conversational memory of LLM agents. ACL 2024, 13851–13870.
- [6] Wu et al. (2024). LongMemEval: Benchmarking chat assistants on long-term interactive memory. arXiv:2410.10813.
- [7] Chhikara et al. (2025). MemO: Building production-ready AI agents with scalable long-term memory. arXiv:2504.19413.
- [8] Rasmussen et al. (2025). Zep: A temporal knowledge graph architecture for agent memory. arXiv:2501.13956.

Acknowledgements

Supervised by Dr Omer Gunes and Prof Philip Treleaven. Work carried out with Otractor, Oxford University Innovation.



Otractor | ziadayed.com

