

Université Cadi Ayyad
Faculté des Sciences Semlalia
Département : Informatique
Master 2

RecLM-RAG

Revolutionizing Recommendations with AI Magic
Zero-Shot, Green, and Crystal Clear!

Réalisé par :

Assoumana Souley Hadiza
Assia Bouamir

Encadré par :

Pr Yassine Afoudi

Présenté devant :

Pr Mohamed Amine Chadi
Pr Mohammed Ameksa
Pr Abdelouafi Ikidid
Pr Yassine Afoudi

Plan:

1

Introduction

2

Etat de l'art

3

Positionnement de RecLM-RAG

4

RecLM-RAG

5

Experiments

6

Demonstration

1

INTRODUCTION

```
130 em.mail{  
131   font-family: 'montserretregular';  
132   background: url('../img/mailico.png') no-repeat center;  
133   display: inline-block;  
134   width: 12px;  
135   height: 14px;  
136   float: left;  
137   margin: 2px 7px 0 0;  
138 }  
139 em.phone{  
140   background: url('../img/phoneico.png') no-repeat center;  
141   display: inline-block;  
142   width: 20px;  
143   height: 18px;  
144   float: left;  
145   margin: 2px 7px 0 0;  
146 }
```

Introduction

Etat de l'art

Positionnement
de RecLM-RAG

RecLM-RAG

Experiments

Demonstration

Contexte & Motivation

SYSTÈMES DE RECOMMANDATION : LA FORCE INVISIBLE DU E-COMMERCE



INDUSTRIAL SOLUTIONS:
80% USE OLD METHODS



EFFICIENT IN LAB,
OUTDATED IN REALITY



Introduction

Etat de l'art

Positionnement
de RecLM-RAG

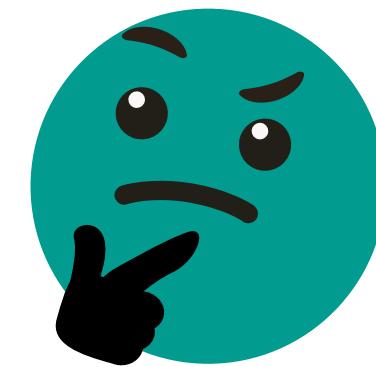
RecLM-RAG

Experiments

Demonstration

Problèmes clés identifiés :

- Interactions utilisateurs éparses
- Catalogues en constante évolution
- Cold-start : nouveaux utilisateurs et produits
- Manque de transparence et d'explicabilité



Pour dépasser ces limites :



- Adopter des modèles sémantiques avancés
- Intégrer des LLMs et RAG
- Prioriser l'éthique et la durabilité
- Evaluer sur des scénarios réalistes

2

Etat de l'art

```
131 }  
132 em.mail{  
133   background: url(..../img/mailico.png) no-repeat center;  
134   display: inline-block;  
135   width: 12px;  
136   height: 14px;  
137   float: left;  
138   margin: 2px 7px 0 0;  
139 }  
140 em.phone{  
141   background: url(..../img/phoneico.png) no-repeat center;  
142   display: inline-block;  
143   width: 20px;  
144   height: 18px;  
145   float: left;  
146   margin: 2px 7px 0 0;  
147 }  
148 em.folder{  
149   background: url(..../img/folderico.png) no-repeat center;  
150   display: inline-block;  
151   width: 16px;  
152   height: 18px;  
153   float: left;  
154   margin: 2px 7px 0 0;  
155 }  
156 em.image{  
157   background: url(..../img/imageico.png) no-repeat center;  
158   display: inline-block;  
159   width: 16px;  
160   height: 18px;  
161   float: left;  
162   margin: 2px 7px 0 0;  
163 }  
164 em.video{  
165   background: url(..../img/videoico.png) no-repeat center;  
166   display: inline-block;  
167   width: 16px;  
168   height: 18px;  
169   float: left;  
170   margin: 2px 7px 0 0;  
171 }  
172 em.audio{  
173   background: url(..../img/audioico.png) no-repeat center;  
174   display: inline-block;  
175   width: 16px;  
176   height: 18px;  
177   float: left;  
178   margin: 2px 7px 0 0;  
179 }  
180 em.pdf{  
181   background: url(..../img/pdfico.png) no-repeat center;  
182   display: inline-block;  
183   width: 16px;  
184   height: 18px;  
185   float: left;  
186   margin: 2px 7px 0 0;  
187 }  
188 em.xls{  
189   background: url(..../img/xlsico.png) no-repeat center;  
190   display: inline-block;  
191   width: 16px;  
192   height: 18px;  
193   float: left;  
194   margin: 2px 7px 0 0;  
195 }  
196 em.ppt{  
197   background: url(..../img/pptico.png) no-repeat center;  
198   display: inline-block;  
199   width: 16px;  
200   height: 18px;  
201   float: left;  
202   margin: 2px 7px 0 0;  
203 }  
204 em.doc{  
205   background: url(..../img/docico.png) no-repeat center;  
206   display: inline-block;  
207   width: 16px;  
208   height: 18px;  
209   float: left;  
210   margin: 2px 7px 0 0;  
211 }  
212 em.txt{  
213   background: url(..../img/txtico.png) no-repeat center;  
214   display: inline-block;  
215   width: 16px;  
216   height: 18px;  
217   float: left;  
218   margin: 2px 7px 0 0;  
219 }  
220 em.html{  
221   background: url(..../img/htmlico.png) no-repeat center;  
222   display: inline-block;  
223   width: 16px;  
224   height: 18px;  
225   float: left;  
226   margin: 2px 7px 0 0;  
227 }  
228 em.zip{  
229   background: url(..../img/zipico.png) no-repeat center;  
230   display: inline-block;  
231   width: 16px;  
232   height: 18px;  
233   float: left;  
234   margin: 2px 7px 0 0;  
235 }  
236 em.rar{  
237   background: url(..../img/rarico.png) no-repeat center;  
238   display: inline-block;  
239   width: 16px;  
240   height: 18px;  
241   float: left;  
242   margin: 2px 7px 0 0;  
243 }  
244 em.mp3{  
245   background: url(..../img/mp3ico.png) no-repeat center;  
246   display: inline-block;  
247   width: 16px;  
248   height: 18px;  
249   float: left;  
250   margin: 2px 7px 0 0;  
251 }  
252 em.jpg{  
253   background: url(..../img/jpgico.png) no-repeat center;  
254   display: inline-block;  
255   width: 16px;  
256   height: 18px;  
257   float: left;  
258   margin: 2px 7px 0 0;  
259 }  
260 em.gif{  
261   background: url(..../img/gifico.png) no-repeat center;  
262   display: inline-block;  
263   width: 16px;  
264   height: 18px;  
265   float: left;  
266   margin: 2px 7px 0 0;  
267 }  
268 em.png{  
269   background: url(..../img/pngico.png) no-repeat center;  
270   display: inline-block;  
271   width: 16px;  
272   height: 18px;  
273   float: left;  
274   margin: 2px 7px 0 0;  
275 }  
276 em.svg{  
277   background: url(..../img/svgico.png) no-repeat center;  
278   display: inline-block;  
279   width: 16px;  
280   height: 18px;  
281   float: left;  
282   margin: 2px 7px 0 0;  
283 }  
284 em.jpeg{  
285   background: url(..../img/jpegico.png) no-repeat center;  
286   display: inline-block;  
287   width: 16px;  
288   height: 18px;  
289   float: left;  
290   margin: 2px 7px 0 0;  
291 }  
292 em.pdf{  
293   background: url(..../img/pdfico.png) no-repeat center;  
294   display: inline-block;  
295   width: 16px;  
296   height: 18px;  
297   float: left;  
298   margin: 2px 7px 0 0;  
299 }  
300 em.xls{  
301   background: url(..../img/xlsico.png) no-repeat center;  
302   display: inline-block;  
303   width: 16px;  
304   height: 18px;  
305   float: left;  
306   margin: 2px 7px 0 0;  
307 }  
308 em.ppt{  
309   background: url(..../img/pptico.png) no-repeat center;  
310   display: inline-block;  
311   width: 16px;  
312   height: 18px;  
313   float: left;  
314   margin: 2px 7px 0 0;  
315 }  
316 em.doc{  
317   background: url(..../img/docico.png) no-repeat center;  
318   display: inline-block;  
319   width: 16px;  
320   height: 18px;  
321   float: left;  
322   margin: 2px 7px 0 0;  
323 }  
324 em.txt{  
325   background: url(..../img/txtico.png) no-repeat center;  
326   display: inline-block;  
327   width: 16px;  
328   height: 18px;  
329   float: left;  
330   margin: 2px 7px 0 0;  
331 }  
332 em.html{  
333   background: url(..../img/htmlico.png) no-repeat center;  
334   display: inline-block;  
335   width: 16px;  
336   height: 18px;  
337   float: left;  
338   margin: 2px 7px 0 0;  
339 }  
340 em.zip{  
341   background: url(..../img/zipico.png) no-repeat center;  
342   display: inline-block;  
343   width: 16px;  
344   height: 18px;  
345   float: left;  
346   margin: 2px 7px 0 0;  
347 }  
348 em.rar{  
349   background: url(..../img/rarico.png) no-repeat center;  
350   display: inline-block;  
351   width: 16px;  
352   height: 18px;  
353   float: left;  
354   margin: 2px 7px 0 0;  
355 }  
356 em.mp3{  
357   background: url(..../img/mp3ico.png) no-repeat center;  
358   display: inline-block;  
359   width: 16px;  
360   height: 18px;  
361   float: left;  
362   margin: 2px 7px 0 0;  
363 }  
364 em.jpg{  
365   background: url(..../img/jpgico.png) no-repeat center;  
366   display: inline-block;  
367   width: 16px;  
368   height: 18px;  
369   float: left;  
370   margin: 2px 7px 0 0;  
371 }  
372 em.gif{  
373   background: url(..../img/gifico.png) no-repeat center;  
374   display: inline-block;  
375   width: 16px;  
376   height: 18px;  
377   float: left;  
378   margin: 2px 7px 0 0;  
379 }  
380 em.png{  
381   background: url(..../img/pngico.png) no-repeat center;  
382   display: inline-block;  
383   width: 16px;  
384   height: 18px;  
385   float: left;  
386   margin: 2px 7px 0 0;  
387 }  
388 em.jpeg{  
389   background: url(..../img/jpegico.png) no-repeat center;  
390   display: inline-block;  
391   width: 16px;  
392   height: 18px;  
393   float: left;  
394   margin: 2px 7px 0 0;  
395 }  
396 em.pdf{  
397   background: url(..../img/pdfico.png) no-repeat center;  
398   display: inline-block;  
399   width: 16px;  
400   height: 18px;  
401   float: left;  
402   margin: 2px 7px 0 0;  
403 }  
404 em.xls{  
405   background: url(..../img/xlsico.png) no-repeat center;  
406   display: inline-block;  
407   width: 16px;  
408   height: 18px;  
409   float: left;  
410   margin: 2px 7px 0 0;  
411 }  
412 em.ppt{  
413   background: url(..../img/pptico.png) no-repeat center;  
414   display: inline-block;  
415   width: 16px;  
416   height: 18px;  
417   float: left;  
418   margin: 2px 7px 0 0;  
419 }  
420 em.doc{  
421   background: url(..../img/docico.png) no-repeat center;  
422   display: inline-block;  
423   width: 16px;  
424   height: 18px;  
425   float: left;  
426   margin: 2px 7px 0 0;  
427 }  
428 em.txt{  
429   background: url(..../img/txtico.png) no-repeat center;  
430   display: inline-block;  
431   width: 16px;  
432   height: 18px;  
433   float: left;  
434   margin: 2px 7px 0 0;  
435 }  
436 em.html{  
437   background: url(..../img/htmlico.png) no-repeat center;  
438   display: inline-block;  
439   width: 16px;  
440   height: 18px;  
441   float: left;  
442   margin: 2px 7px 0 0;  
443 }  
444 em.zip{  
445   background: url(..../img/zipico.png) no-repeat center;  
446   display: inline-block;  
447   width: 16px;  
448   height: 18px;  
449   float: left;  
450   margin: 2px 7px 0 0;  
451 }  
452 em.rar{  
453   background: url(..../img/rarico.png) no-repeat center;  
454   display: inline-block;  
455   width: 16px;  
456   height: 18px;  
457   float: left;  
458   margin: 2px 7px 0 0;  
459 }  
460 em.mp3{  
461   background: url(..../img/mp3ico.png) no-repeat center;  
462   display: inline-block;  
463   width: 16px;  
464   height: 18px;  
465   float: left;  
466   margin: 2px 7px 0 0;  
467 }  
468 em.jpg{  
469   background: url(..../img/jpgico.png) no-repeat center;  
470   display: inline-block;  
471   width: 16px;  
472   height: 18px;  
473   float: left;  
474   margin: 2px 7px 0 0;  
475 }  
476 em.gif{  
477   background: url(..../img/gifico.png) no-repeat center;  
478   display: inline-block;  
479   width: 16px;  
480   height: 18px;  
481   float: left;  
482   margin: 2px 7px 0 0;  
483 }  
484 em.png{  
485   background: url(..../img/pngico.png) no-repeat center;  
486   display: inline-block;  
487   width: 16px;  
488   height: 18px;  
489   float: left;  
490   margin: 2px 7px 0 0;  
491 }  
492 em.jpeg{  
493   background: url(..../img/jpegico.png) no-repeat center;  
494   display: inline-block;  
495   width: 16px;  
496   height: 18px;  
497   float: left;  
498   margin: 2px 7px 0 0;  
499 }  
500 em.pdf{  
501   background: url(..../img/pdfico.png) no-repeat center;  
502   display: inline-block;  
503   width: 16px;  
504   height: 18px;  
505   float: left;  
506   margin: 2px 7px 0 0;  
507 }  
508 em.xls{  
509   background: url(..../img/xlsico.png) no-repeat center;  
510   display: inline-block;  
511   width: 16px;  
512   height: 18px;  
513   float: left;  
514   margin: 2px 7px 0 0;  
515 }  
516 em.ppt{  
517   background: url(..../img/pptico.png) no-repeat center;  
518   display: inline-block;  
519   width: 16px;  
520   height: 18px;  
521   float: left;  
522   margin: 2px 7px 0 0;  
523 }  
524 em.doc{  
525   background: url(..../img/docico.png) no-repeat center;  
526   display: inline-block;  
527   width: 16px;  
528   height: 18px;  
529   float: left;  
530   margin: 2px 7px 0 0;  
531 }  
532 em.txt{  
533   background: url(..../img/txtico.png) no-repeat center;  
534   display: inline-block;  
535   width: 16px;  
536   height: 18px;  
537   float: left;  
538   margin: 2px 7px 0 0;  
539 }  
540 em.html{  
541   background: url(..../img/htmlico.png) no-repeat center;  
542   display: inline-block;  
543   width: 16px;  
544   height: 18px;  
545   float: left;  
546   margin: 2px 7px 0 0;  
547 }  
548 em.zip{  
549   background: url(..../img/zipico.png) no-repeat center;  
550   display: inline-block;  
551   width: 16px;  
552   height: 18px;  
553   float: left;  
554   margin: 2px 7px 0 0;  
555 }  
556 em.rar{  
557   background: url(..../img/rarico.png) no-repeat center;  
558   display: inline-block;  
559   width: 16px;  
560   height: 18px;  
561   float: left;  
562   margin: 2px 7px 0 0;  
563 }  
564 em.mp3{  
565   background: url(..../img/mp3ico.png) no-repeat center;  
566   display: inline-block;  
567   width: 16px;  
568   height: 18px;  
569   float: left;  
570   margin: 2px 7px 0 0;  
571 }  
572 em.jpg{  
573   background: url(..../img/jpgico.png) no-repeat center;  
574   display: inline-block;  
575   width: 16px;  
576   height: 18px;  
577   float: left;  
578   margin: 2px 7px 0 0;  
579 }  
580 em.gif{  
581   background: url(..../img/gifico.png) no-repeat center;  
582   display: inline-block;  
583   width: 16px;  
584   height: 18px;  
585   float: left;  
586   margin: 2px 7px 0 0;  
587 }  
588 em.png{  
589   background: url(..../img/pngico.png) no-repeat center;  
590   display: inline-block;  
591   width: 16px;  
592   height: 18px;  
593   float: left;  
594   margin: 2px 7px 0 0;  
595 }  
596 em.jpeg{  
597   background: url(..../img/jpegico.png) no-repeat center;  
598   display: inline-block;  
599   width: 16px;  
600   height: 18px;  
601   float: left;  
602   margin: 2px 7px 0 0;  
603 }  
604 em.pdf{  
605   background: url(..../img/pdfico.png) no-repeat center;  
606   display: inline-block;  
607   width: 16px;  
608   height: 18px;  
609   float: left;  
610   margin: 2px 7px 0 0;  
611 }  
612 em.xls{  
613   background: url(..../img/xlsico.png) no-repeat center;  
614   display: inline-block;  
615   width: 16px;  
616   height: 18px;  
617   float: left;  
618   margin: 2px 7px 0 0;  
619 }  
620 em.ppt{  
621   background: url(..../img/pptico.png) no-repeat center;  
622   display: inline-block;  
623   width: 16px;  
624   height: 18px;  
625   float: left;  
626   margin: 2px 7px 0 0;  
627 }  
628 em.doc{  
629   background: url(..../img/docico.png) no-repeat center;  
630   display: inline-block;  
631   width: 16px;  
632   height: 18px;  
633   float: left;  
634   margin: 2px 7px 0 0;  
635 }  
636 em.txt{  
637   background: url(..../img/txtico.png) no-repeat center;  
638   display: inline-block;  
639   width: 16px;  
640   height: 18px;  
641   float: left;  
642   margin: 2px 7px 0 0;  
643 }  
644 em.html{  
645   background: url(..../img/htmlico.png) no-repeat center;  
646   display: inline-block;  
647   width: 16px;  
648   height: 18px;  
649   float: left;  
650   margin: 2px 7px 0 0;  
651 }  
652 em.zip{  
653   background: url(..../img/zipico.png) no-repeat center;  
654   display: inline-block;  
655   width: 16px;  
656   height: 18px;  
657   float: left;  
658   margin: 2px 7px 0 0;  
659 }  
660 em.rar{  
661   background: url(..../img/rarico.png) no-repeat center;  
662   display: inline-block;  
663   width: 16px;  
664   height: 18px;  
665   float: left;  
666   margin: 2px 7px 0 0;  
667 }  
668 em.mp3{  
669   background: url(..../img/mp3ico.png) no-repeat center;  
670   display: inline-block;  
671   width: 16px;  
672   height: 18px;  
673   float: left;  
674   margin: 2px 7px 0 0;  
675 }  
676 em.jpg{  
677   background: url(..../img/jpgico.png) no-repeat center;  
678   display: inline-block;  
679   width: 16px;  
680   height: 18px;  
681   float: left;  
682   margin: 2px 7px 0 0;  
683 }  
684 em.gif{  
685   background: url(..../img/gifico.png) no-repeat center;  
686   display: inline-block;  
687   width: 16px;  
688   height: 18px;  
689   float: left;  
690   margin: 2px 7px 0 0;  
691 }  
692 em.png{  
693   background: url(..../img/pngico.png) no-repeat center;  
694   display: inline-block;  
695   width: 16px;  
696   height: 18px;  
697   float: left;  
698   margin: 2px 7px 0 0;  
699 }  
700 em.jpeg{  
701   background: url(..../img/jpegico.png) no-repeat center;  
702   display: inline-block;  
703   width: 16px;  
704   height: 18px;  
705   float: left;  
706   margin: 2px 7px 0 0;  
707 }  
708 em.pdf{  
709   background: url(..../img/pdfico.png) no-repeat center;  
710   display: inline-block;  
711   width: 16px;  
712   height: 18px;  
713   float: left;  
714   margin: 2px 7px 0 0;  
715 }  
716 em.xls{  
717   background: url(..../img/xlsico.png) no-repeat center;  
718   display: inline-block;  
719   width: 16px;  
720   height: 18px;  
721   float: left;  
722   margin: 2px 7px 0 0;  
723 }  
724 em.ppt{  
725   background: url(..../img/pptico.png) no-repeat center;  
726   display: inline-block;  
727   width: 16px;  
72
```

Introduction

Etat de l'art

Positionnement
de RecLM-RAG

RecLM-RAG

Experiments

Demonstration

Collaborative Filtering & Deep Models

Collaborative Filtering (CF) domine l'industrie : Netflix, YouTube, Amazon

Matrix Factorization : SVD, SVD++, ALS → scalable et robuste sur données denses

Neighborhood-based : User-KNN, Item-KNN → efficace sur sparsité mais problème cold-start

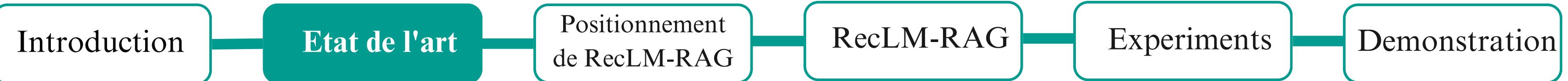
Deep Learning :

Sequential : GRU4Rec, SASRec, BERT4Rec, TiSASRec

Graph-based : LightGCN, NGCF, PinSage

Limites :

- Besoin de millions d'interactions
- Pas d'explications en langage naturel
- Ignorent l'éthique et l'impact environnemental



LLM-based Recommendation

Prompting-only

Exemples : P5, TALLRec, ZeroLLM, ChatGPT

Avantages : zéro-shot, unification des tâches text-to-text

Limites : long-tail items, contrôle éthique limité

Parameter-efficient fine-tuning

Exemples : LLM-Rec, RecLLM, CTRL

Avantages : SOTA performance

Limites : coût d'entraînement élevé, perte de zéro-shot

Conversational / interactive

Exemples : Chat-Rec, CRSRec, RecInDial

Génération d'explications naturelles

Limites : dépendance aux modèles CF → cold-start & biais

Introduction

Etat de l'art

Positionnement
de RecLM-RAG

RecLM-RAG

Experiments

Demonstration

Retrieval-Augmented Generation (RAG)

Combinaison dense retrieval + LLMs :

- RLMRec, E-RAG → items via encoders + re-ranking
- RecRanker, RankRAG → ranking listwise
- Self-RAG, RAG-Fusion → query rewriting, amélioration retrieval

Extensions multimodales : CLIP4Rec, M6-Rec

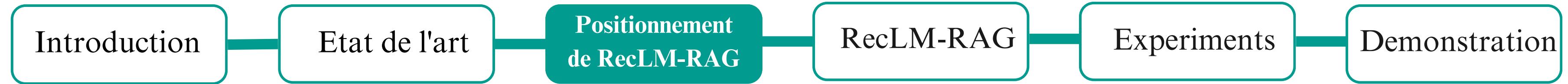
Limites :

- Pas de zéro-shot complet
- Pas d'intégration directe de durabilité
- Dépendances aux CF

3

Positionnement de RecLM-RAG

```
131 em.mail{  
132   background: url(..../img/mailico.png) no-repeat center;  
133   display: inline-block;  
134   width: 12px;  
135   height: 14px;  
136   float: left;  
137   margin: 2px 7px 0 0;  
138 }  
139 em.phone{  
140   background: url(..../img/phonico.png) no-repeat center;  
141   display: inline-block;  
142   width: 20px;  
143   height: 18px;  
144   float: left;  
145   margin: 2px 7px 0 0;  
146 }
```



RecLM-RAG vs State-of-the-Art

Fonctionnalité	CF-based	LLM-based	RAG	RecLM-RAG
Zéro-shot	✗	partiel	partiel	✓
Explicabilité NL	✗	✓	✓	✓
Durabilité intégrée	✗	✗	✗	✓
Indépendance CF	✗	✗	✗	✓

4

RecLM-RAG: Overview

Introduction

Etat de l'art

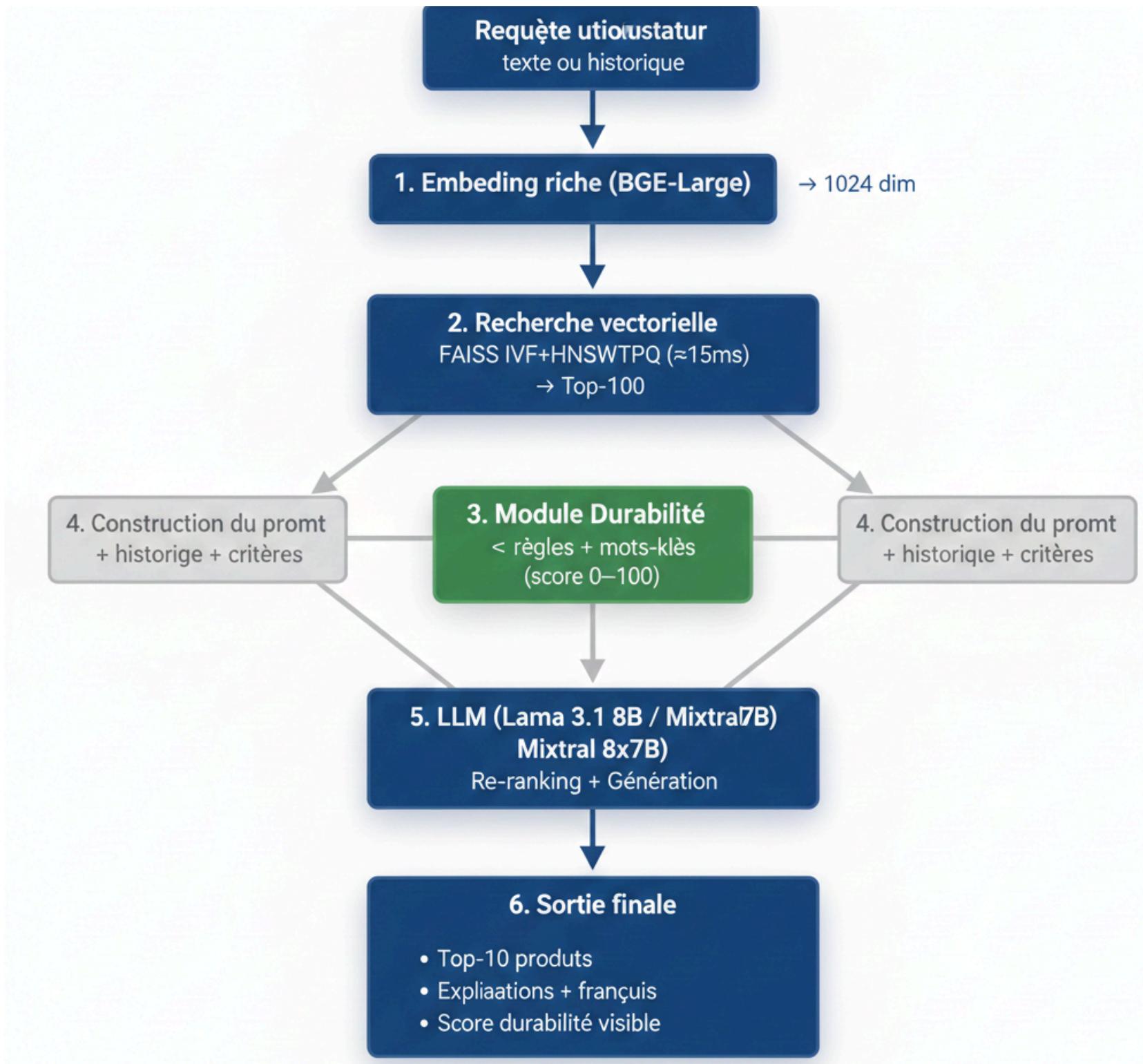
Positionnement
de RecLM-RAG

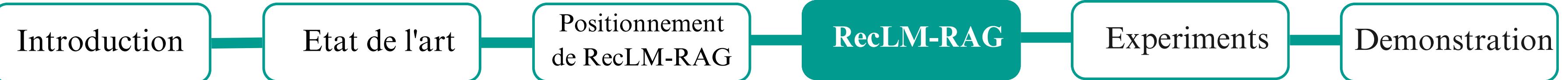
RecLM-RAG

Experiments

Demonstration

RecLM-RAG: Zero-Shot, End-to-End Recommendation





Entrée Utilisateur

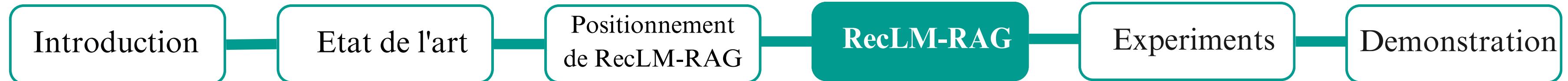
Deux modes d'entrée :

- **Query mode** : texte naturel
- **History mode** : “User bought: {liste d’items}”



Encodage via BGE-Large-v1.5 (1024 dims)

Support multilingue



Step 2: FAISS Vector Index

- **Division de l'espace en clusters**

Les vecteurs sont regroupés en clusters (IVF) pour limiter la recherche à une zone pertinente.

- **Exploration hiérarchique**

FAISS explore d'abord les clusters les plus proches avant d'affiner la recherche.

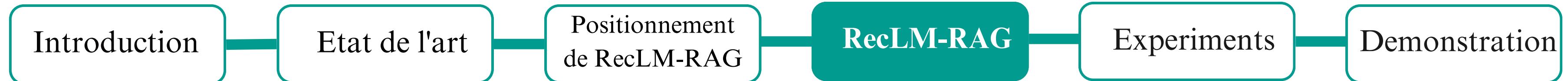
- **Graphe navigable Small World (HNSW)**

Un graphe de voisinage permet de naviguer rapidement entre vecteurs proches.

- **Compression des vecteurs**

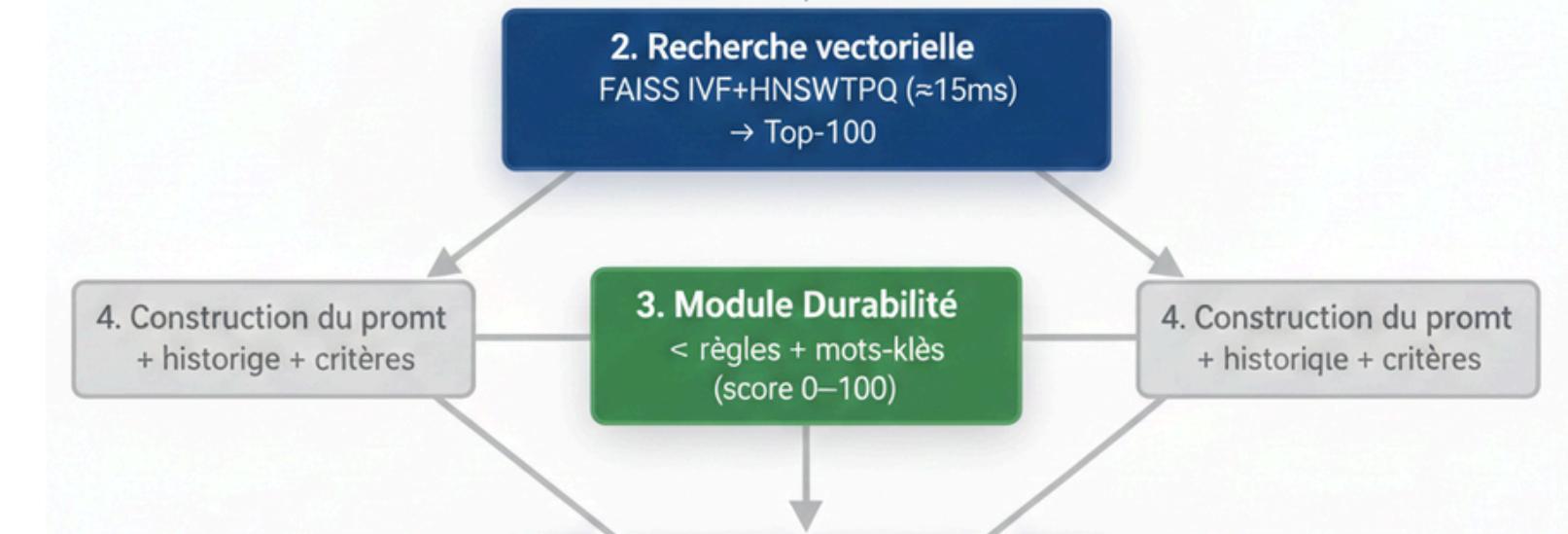
La quantification (PQ) réduit la mémoire et accélère la recherche.

2. Recherche vectorielle
FAISS IVF+HNSWTPQ ($\approx 15\text{ms}$)
→ Top-100



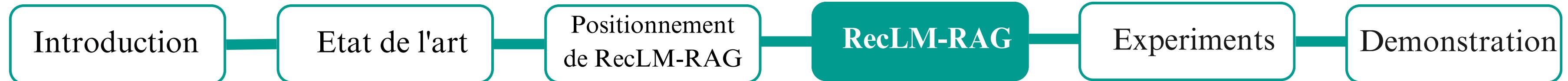
Step 3: Ethical & Sustainability Module

- Score $S(i) \in [0,100]$ basé sur mots-clés
 - Positif : recycled, organic, fair trade...
 - Négatif : virgin plastic, fast fashion...



$$S(i) = 100 \cdot (w_{\text{pos}} \cdot I[\text{positive keywords}] - w_{\text{neg}} \cdot I[\text{negative keywords}] + b_{\text{cert}})$$

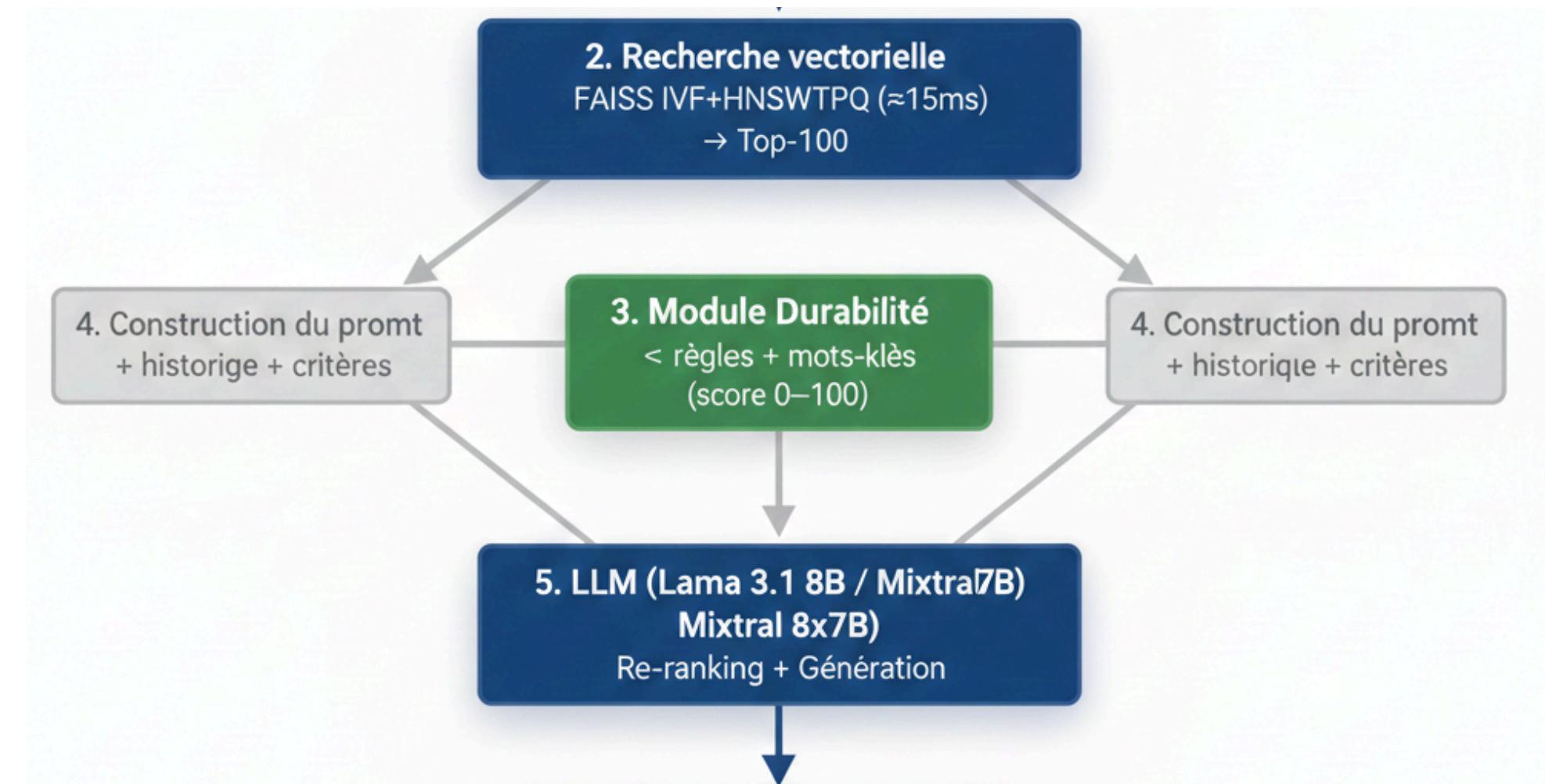
- Temps d'exécution : <1 ms
- Filtrage éthique explicite des produits



Step 4: Prompt Construction

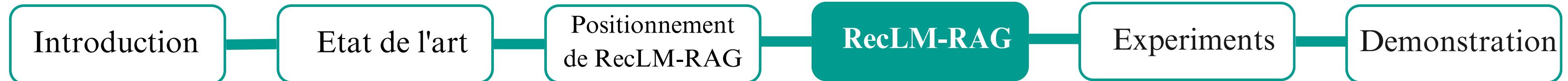
Fusion de :

- requête utilisateur
- historique d'achat
- top-100 produits (FAISS)
- score de durabilité
- Scoring Hybride (30% Intelligence Collective +70% Sémantique)



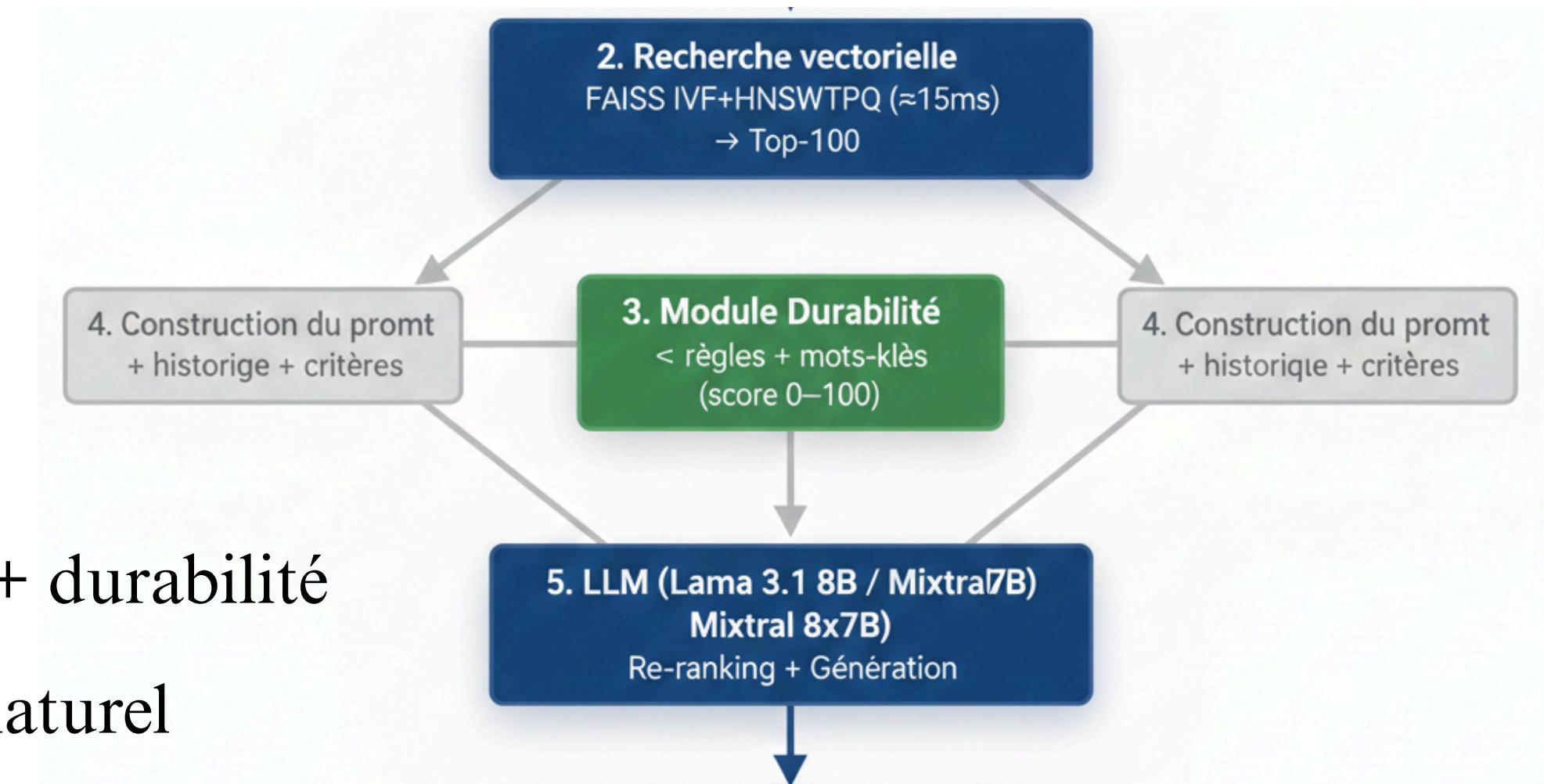
Format textuel unique pour le LLM

Objectif : classement + explication



steep 5 – LLM Reasoning

- Modèle : LLaMA 3.1 / Mixtral
- Le LLM :
 - classe les produits selon pertinence + durabilité
 - génère des explications en langage naturel
 - retourne un JSON structuré



Introduction

Etat de l'art

Positionnement
de RecLM-RAG

RecLM-RAG

Experiments

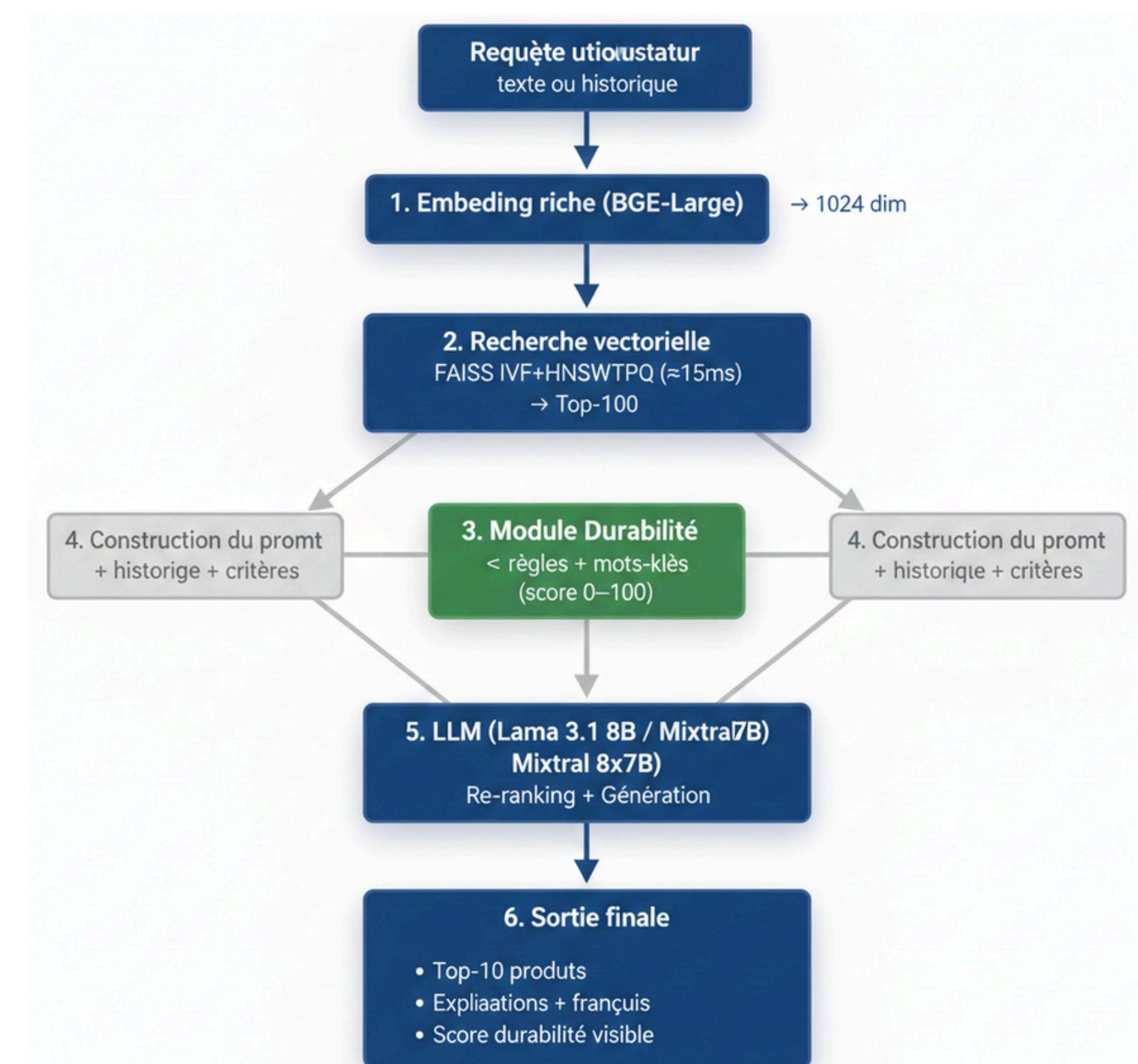
Demonstration

RecLM-RAG: Zero-Shot, End-to-End Recommendation

Step 5 & 6: Algorithm & Complexity

RecLM-RAG Inference :

1. Encoder requête/historique
2. Récupérer 100 items FAISS
3. Ajouter score durabilité
4. Construire prompt
5. Lancer LLM pour re-ranking + explications
6. Extraire Top-10 recommandations



5

Experiments

Introduction

Etat de l'art

Positionnement
de RecLM-RAG

RecLM-RAG

Experiments

Demonstration

Datasets & Experimental Protocol

Datasets : Amazon Reviews 2018

Électronique : produits techniques, specs détaillées

Vêtements & accessoires : préférences subjectives

Cold-start : 10k utilisateurs sans historique → reliance sur query/metadata seulement

Statistiques principales :

	Electronics	Clothing, Shoes & Jewelry
# Users	192,403	278,154
# Items	63,001	85,547
# Interactions	7,824,482	5,748,292
Sparsity	99.94%	99.88%
Avg. review length	78 tokens	92 tokens
Avg. items per user	40.7	20.7
Avg. users per item	124.2	67.2



Baselines

- **Popularity (non-personnalisé)** : recommande les items les plus populaires, identiques pour tous les utilisateurs.
- **SVD (Matrix Factorization)** : décompose la matrice user–item pour apprendre des facteurs latents.
- **LightGCN** : modèle graphe qui propage les interactions user–item sans couches complexes.
- **SASRec** : transformer séquentiel qui prédit le prochain item à partir de l'historique.
- **P5** : modèle T5 adapté à la recommandation en few-shot via du texte.
- **Deep Learning setup** : entraînement avec PyTorch 2.1 sur GPU, batch=256, optimiseur AdamW, hyperparamètres via grid search.

Introduction

Etat de l'art

Positionnement
de RecLM-RAG

RecLM-RAG

Experiments

Demonstration

Overall Performance

Méthode	Recall@10	NDCG@10	Diversity@10	Long-Tail (%)
Popularity	0.189	0.312	0.21	4.2
SVD	0.256	0.378	0.29	8.7
LightGCN	0.341	0.456	0.38	14.3
SASRec	0.367	0.471	0.41	16.8
P5	0.398	0.489	0.47	19.4
RecLM-RAG (Llama)	0.498 [†]	0.562 [†]	0.71 [†]	38.4 [†]
RecLM-RAG (Mixtral)	0.514 [†]	0.579 [†]	0.73 [†]	41.2 [†]

Gains massifs sur précision, diversité, long-tail et cold-start

Recall@10 → "Combien de bons produits j'ai réussi à proposer dans les 10 ?"

NDCG@10 → "Les meilleurs produits sont-ils bien placés en haut de liste ?"

Diversity@10 → "Mes 10 recommandations sont-elles variées ou toutes pareilles ?"

Long-Tail (%) → "Je recommande surtout des best-sellers ou aussi des produits de niche ?"



Ablation Studyt

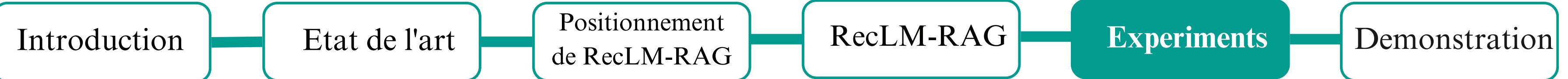
Variante	NDCG@10	Diversity@10	Impact
Full RecLM-RAG	0.562	0.71	-
sans module durabilité	0.548	0.68	-2.5% / -4.2%
sans historique utilisateur	0.519	0.70	-7.7% / -1.4%
Top-50 candidates	0.541	0.69	-3.7% / -2.8%
BGE-base vs BGE-large	0.512	0.66	-8.9% / -7.0%
FAISS only	0.472	0.62	-16% / -12.7%

Importance : durabilité, historique, embeddings avancés, LLM re-ranking

6

Demonstration

```
        font-size: 1em;
        font-family: 'montserratregular';
    }
}
131
132 em.mail{
133     background: url(../img/mailico.png) no-repeat center;
134     display: inline-block;
135     width: 12px;
136     height: 14px;
137     float: left;
138     margin: 2px 7px 0 0;
139 }
140 em.phone{
141     background: url(../img/phoneico.png) no-repeat center;
142     display: inline-block;
143     width: 20px;
144     height: 18px;
145     float: left;
146     margin: 3px 8px 0 0;
147 }
```



Future work

- Multimodalité : intégrer CLIP/LLaVA pour récupérer à la fois texte et images.
- Module de durabilité appris : entraîner sur datasets certifiés (ex. Higg Index, labels FairTrade) pour plus de précision.
- Optimisation LLM : spéculative + quantification 3-bit pour latence <500 ms.
- Auto-reflection : tokens Self-RAG pour critique et re-ranking automatiques.
- Tests à grande échelle : A/B testing sur un partenaire e-commerce réel pour validation terrain.



Merci pour votre attention
