OneSearch-V2: A Latent Reasoning Enhanced Self-distillation Generative Search Framework

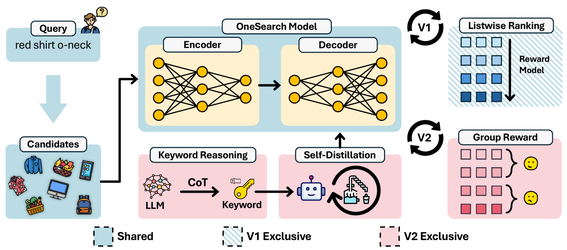
Ben Chen[1,2], Siyuan Wang[1], Yufei Ma[1], Zihan Liang[1], Han Li, Chenyi Lei, Wenwu Ou, Kun Gai

Kuaishou Technology [1] Equal contribution [2] Corresponding author benchen4395@gmail.com

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Generative retrieval predicts item semantic IDs (SIDs). Complex queries and user context require richer intent reasoning, while online search must keep serving latency low.



V2 internalizes reasoning and replaces the separate reward model with direct behavior feedback.

Four-stage training

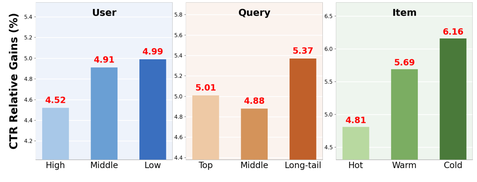
1. Semantic alignment with keyword-based CoT tasks.
2. Query–item co-occurrence modeling.
3. User personalization with self-distillation, R-Drop, FGM and focal loss.
4. Listwise DPO followed by TPMA-based GRPO.

Keyword-based CoT ablation

Integration	Head HR@10	Tail HR@10
Baseline	0.2362	0.1952
+ CoT tasks	0.2419	0.1963
+ direct CoT generation	0.1116	0.0809
+ RAG input augmentation	**0.2438**	**0.1973**

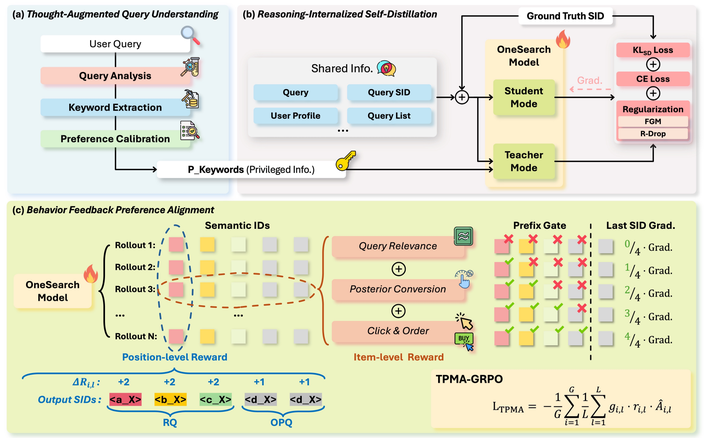
The last two rows are alternative SFT Stage 3 additions after CoT tasks. Direct CoT generation substantially lowers HR@10.

CTR gains across traffic segments



Most recent A/B run, relative gains over V1. This breakdown uses a different aggregation from the overall online results.

OneSearch-V2 framework



Compact keyword reasoning

Offline query analysis, keyword extraction and preference calibration create privileged inputs for the teacher.

Shared-weight self-distillation

The teacher sees keyword-augmented inputs. The student learns its distributions from ordinary inputs, with R-Drop and FGM for consistency and robustness.

Position-aware preference alignment

TPMA assigns credit to hierarchical SID positions using relevance, posterior CTR and click/order rewards.

Self-distillation alternatives

Method	Order HR@10	Click HR@10
Base (S), no keywords	0.2094	0.2266
Base (T), with keywords	0.2139	0.2327
Special-token distillation	0.2092	0.2335
Latent + CODI	0.2105	0.2269
EMA-mode	0.2097	0.2317
Joint-mode	0.2156	0.2348
Self-Distill (T)	0.2155	0.2397
Self-Distill (S), ours	**0.2163**	**0.2398**

S: no keyword input at inference. T: keyword input. The self-distilled student retains the original serving architecture.

Offline component contributions

Order: 7,229 samples. Click: 30K. Higher is better.

Configuration	Order HR@10	Order MRR@10	Click HR@10	Click MRR@10
V1 baseline	0.2046	0.0985	0.2231	0.0728
+ CoT tasks	0.2094	0.1008	0.2266	0.0731
+ self-distillation	0.2163	0.1017	0.2398	0.0757
+ R-Drop	0.2168	0.1045	0.2398	0.0760
+ FGM	0.2180	0.1047	0.2422	0.0766
+ focal loss (SFT full)	0.2214	0.1048	0.2471	0.0788
SFT full + PARS	0.2221	0.1067	0.2538	0.0809
SFT full + GRPO	0.2248	0.1106	0.2481	0.0798
SFT full + TPMA	0.2265	0.1136	0.2498	0.0815
V2 (DPO + TPMA)	**0.2314**	**0.1151**	**0.2568**	**0.0833**

Rows through focal loss are cumulative. PARS, GRPO and TPMA are separate additions to SFT full. V2 adds DPO + TPMA.

Online gains by component

Relative gains over V1. All A/B entries have $p < 0.05$.

Metric	V2 RAG	V2 Reason	V2 full
Item CTR	+0.52%	+2.59%	**+3.98%**
PV CTR	+0.77%	+1.42%	**+1.17%**
PV CVR	+0.63%	+2.21%	**+2.90%**
Buyer volume	+1.04%	+1.50%	**+2.07%**
Order volume	+1.07%	+1.57%	**+2.11%**

RAG: CoT tasks. Reason: + self-distillation. Full: + behavior alignment. Full improves conversion with a smaller PV CTR lift.

Human evaluation

200 queries, 3,200 query-item pairs. Relative gains over V1.

Metric	V2 Reason	V2 full
Page good rate	+1.12%	**+1.37%**
Item quality	+0.28%	**+0.55%**
Query-item relevance	+1.01%	**+1.65%**

Scope: one commercial platform and short-term online evaluation.