SkyDistill : Navigating Fuzzy Flight Search Ranking via Precise Intent Distillation.

Maolei Huang, Shuhan Song *, Huawei Cao
Fliggy (Alibaba Group), ICT (UCAS)
hhangmaolei.hml@alibaba-inc.com , songshuhan@ict.ac.cn

SIGIR2026
MELBOURNE | NAARM
AUSTRALIA

1. Background

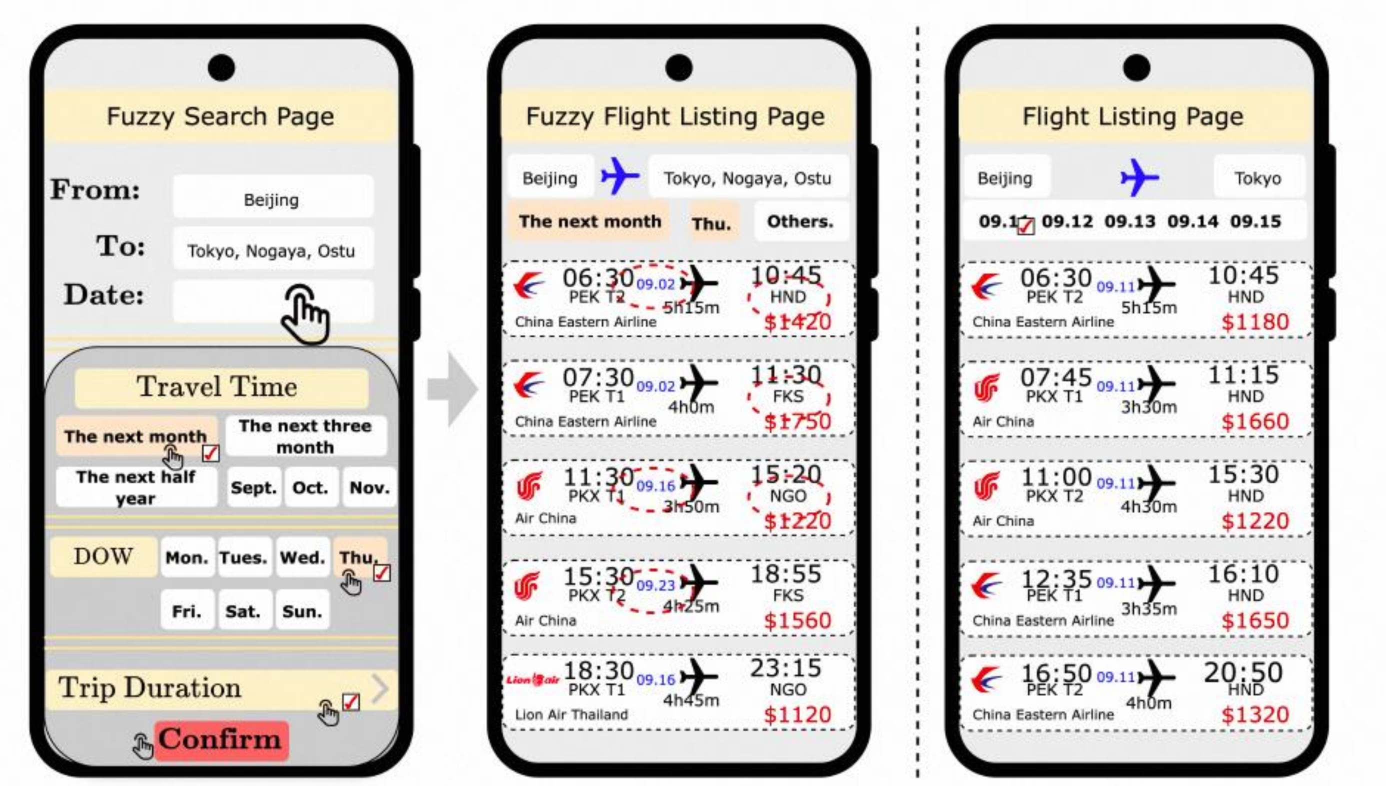


Figure 1: Comparison of Fuzzy vs. Precise Flight Search. (Left) Fuzzy input allows multi-destination and flexible temporal queries. (Middle) Fuzzy results present heterogeneous itineraries across varied dates and airports. (Right) Traditional precise search follows a rigid one-to-one mapping.

a. **Strategic Importance:** Fuzzy Flight Search has become a critical gateway for large-scale platforms like Fliggy, processing high-frequency requests from over ***200,000*** daily active users (DAUs). It serves as a primary hub for users seeking discovery rather than specific bookings.

b. **Shift in Paradigm:** The search interaction logic has fundamentally evolved from ***Constraint-based Retrieval*** to ***Exploratory Discovery***. Users now frequently employ broad, flexible queries, such as "any Thursday next month" or "multi-destination options", prioritizing travel inspiration over scheduling.

c. **Decision-Making Manifold:** This environment requires the system to act as a sophisticated recommendation engine that navigates high-dimensional decision spaces.

2. Challenge

a. **Sparse Signals & Low Conversion:** Unlike precise search, where intent is explicit, fuzzy queries lack clear signals. The inherent ambiguity in user requirements complicates the ranking process, manifesting in low conversion rates ***(1.6% UV-CVR)*** and significant difficulty in satisfying stochastic user intentions.

b. **Domain Discrepancy & Misalignment:** Standard ranking models trained on deterministic, precise-search data struggle with "***distributional shift***" when applied to fuzzy scenarios. This leads to a persistent performance gap, where superior offline metrics (e.g., NDCG, AUC) fail to translate into meaningful online conversion improvements..

c. **Noise Propagation in Knowledge Transfer:** Simply applying knowledge distillation is insufficient due to the inherent noise in fuzzy search data. The system faces the critical challenge of preventing "***dark knowledge***" from contaminating the student model, requiring a nuanced, uncertainty-aware approach to effectively align feature spaces between heterogeneous domains.

3. Methodology (MCKD)

3.1 Framework Overview

The framework bridges the domain gap between precise and fuzzy search. It comprises two stages:

Teacher Pre-training: A high-capacity teacher $\mathcal{T}$ encodes complex user-item interactions within the feature-rich precise search domain $\mathcal{D}_s$.

Distillation-guided Learning: The student $\mathcal{S}$ is trained on restricted-feature fuzzy queries $\mathcal{D}_t$, regularized by multi-level latent knowledge extracted from the teacher to resolve signal sparsity.

3.2 Dual-Objective Logit Distillation

We employ temperature scaling to amplify "***dark knowledge***" in the probability distribution. The student learns from both soft teacher predictions and hard ground-truth labels, ensuring a balance between teacher-mimicry and task-specific accuracy.

$$p_i^{\mathcal{T}} = \frac{1}{1+\exp(-z_i^{\mathcal{T}}/T)}$$

$$\mathcal{L}_{base} = \lambda \cdot \mathcal{L}_{soft}(p_i^{\mathcal{T}}, q_i^{\mathcal{S}}) + (1-\lambda)\cdot \mathcal{L}_{hard}(y_i, z_i^{\mathcal{S}})$$

$$\mathcal{L}_{soft} = \mathcal{L}_{log}(p_i^{\mathcal{T}}, q_i^{\mathcal{S}}) + \mathcal{L}_{kl}(p_i^{\mathcal{T}}, q_i^{\mathcal{S}}),\ \mathcal{L}_{hard} = \mathcal{L}_{ce}(y_i, z_i^{\mathcal{S}})$$

3.3 Feature-Level Hint Learning

Logit distillation alone is insufficient for capturing hierarchical representations. We align the teacher's intermediate latent space with the student's via a learnable linear projection $\mathrm{Proj}(\cdot)$:

$$\mathcal{L}_{hint} = \|\mathbf{h}_m^{\mathbf{T}} - \mathrm{Proj}(\mathbf{h}_n^{\mathbf{S}})\|_2^2$$

This provides a robust inductive bias, forcing the student to inherit the teacher's high-order feature abstraction capabilities.

3.4 Uncertainly-Aware Adaptive Distillation

To prevent "***noise propagation***" from **out-of-distribution** queries, we modulate the distillation weight α_i based on the teacher's predictive entropy.

Confidence Reinforcement: If the teacher is certain $\mathbb{H} \to o$, α_i is maximized to facilitate knowledge transfer.

Uncertainty Suppression: If the teacher is ambiguous $\mathbb{H} \to$ peak , α_i decays exponentially to prevent the student from inheriting unreliable signals.

$$\alpha_i = \alpha_{base} \cdot \exp\left(-\gamma \cdot \frac{\mathbb{H}(p_i^{\mathcal{T}})}{H_{max}}\right)$$

3.5 Pairwise Ranking Distillation

Flight search is a comparative decision process. We force the student to replicate the teacher's preference manifold by matching the relative rankings of candidates(i, j):

$$\mathcal{L}_{pair} = \sum_{(i,j)\in\mathcal{P}} \mathrm{BCE}\left(\sigma(z_i^{\mathcal{S}} - z_j^{\mathcal{S}}), \sigma(z_i^{\mathcal{T}} - z_j^{\mathcal{T}})\right)$$

This ensures that the relative preference orders, crucial for user utility, are preserved even when absolute probability signals are sparse.

3.6 Total Training Objective

The final optimization objective is a multi-task formulation that integrates point-wise accuracy, structural representation, and pair-wise ranking consistency:

$$\mathcal{L}_{total} = \mathcal{L}_{base} + \lambda_1 \mathcal{L}_{hint} + \lambda_2 \mathcal{L}_{pair}$$

Post-training, all teacher components and projection layers are discarded, ensuring a lightweight, high-speed inference pipeline for online production.

4. Offline Evaluation

Table 1: Performance comparison on production datasets (Best values are in bold; next-best values are underlined).

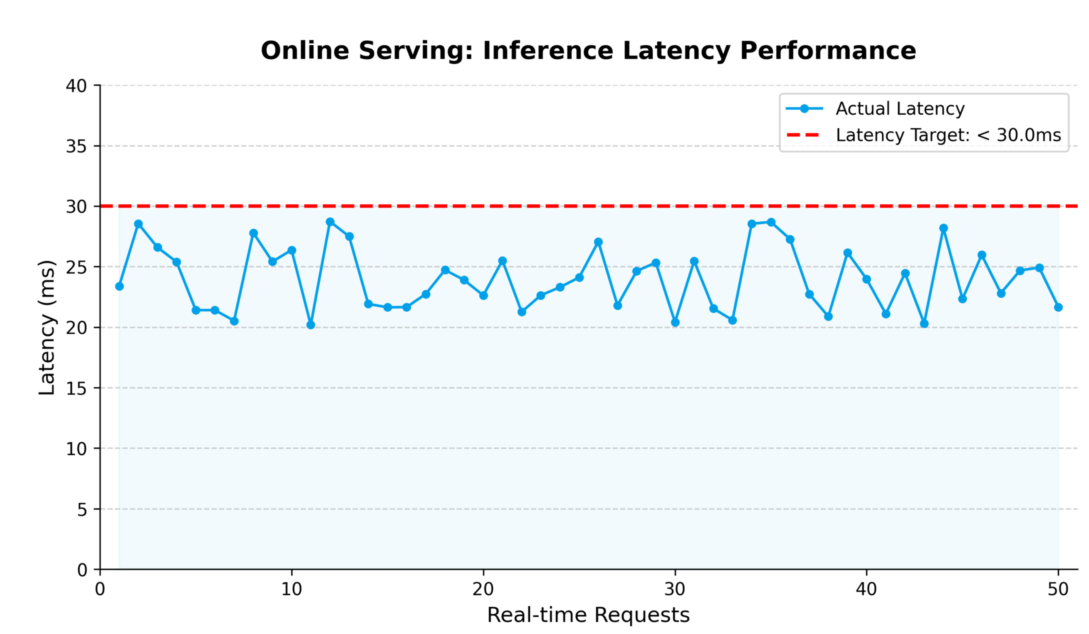
Categories	Methods	HR@1	HR@5	HR@10	HR@15	MRR@1	MRR@5	MRR@10	MRR@15	AUC	GAUC
Rule	Cheapest	48.66%	72.99%	84.06%	92.78%	48.66%	59.17%	60.28%	61.91%	76.02%	81.48%
	Shortest	48.23%	72.78%	83.89%	92.99%	48.23%	59.12%	60.25%	61.88%	76.01%	81.54%
	Policies	47.77%	72.86%	84.67%	93.67%	47.77%	59.36%	60.62%	62.06%	77.09%	82.01%
Single-Task	WDL	48.98%	73.19%	84.84%	93.01%	48.98%	59.01%	60.10%	61.91%	76.37%	81.63%
	DIN	49.99%	73.18%	85.02%	93.88%	49.99%	58.99%	60.11%	61.78%	76.15%	81.82%
	DIEN	50.16%	73.22%	85.11%	92.87%	50.16%	58.91%	60.01%	61.89%	76.73%	81.26%
Multi-Task	MMoE	50.82%	73.87%	85.97%	93.38%	50.82%	59.77%	61.03%	62.05%	76.59%	81.28%
	ESMM	51.36%	73.99%	85.91%	94.21%	51.36%	59.72%	61.16%	62.51%	76.83%	81.24%
	PLE	51.22%	74.03%	85.89%	94.13%	51.22%	59.89%	61.24%	62.55%	76.66%	81.12%
Itinerary	PFRN	50.16%	<u>74.48%</u>	86.36%	94.26%	50.16%	60.66%	63.84%	64.87%	77.15%	81.41%
	DCM	50.11%	74.01%	86.17%	94.06%	50.11%	60.06%	63.31%	64.77%	76.89%	81.59%
	PlanRanker	50.18%	74.12%	86.32%	94.56%	50.18%	60.04%	64.01%	65.03%	77.08%	81.77%
	CPNet	<u>52.06%</u>	73.98%	86.11%	<u>94.77%</u>	<u>52.06%</u>	60.11%	64.02%	65.10%	77.06%	81.87%
	DCRNet	51.73%	74.26%	<u>86.78%</u>	94.18%	51.73%	<u>60.19%</u>	<u>64.16%</u>	<u>65.13%</u>	<u>77.13%</u>	<u>82.01%</u>
	MCKD(ours)	**53.24%**	**75.18%**	**87.97%**	**95.62%**	**53.24%**	**61.33%**	**64.62%**	**65.76%**	**78.86%**	**83.42%**

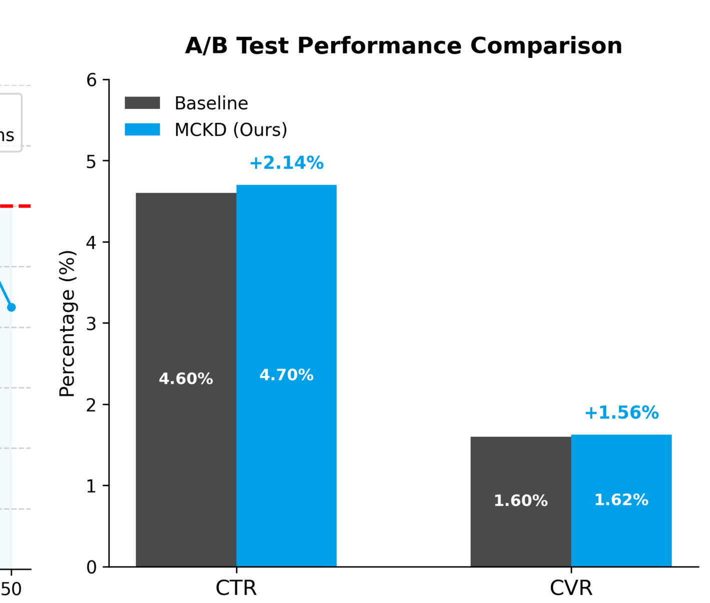
Table 2: Ablation Study of MCKD.

Methods	HR@5	MRR@5	GAUC	AUC
w/o Sec. 2.2	74.26%	60.16%	0.6210	77.11%
w/o Sec. 2.3	74.88%	61.02%	0.6214	78.23%
w/o Sec. 2.4	74.72%	60.26%	0.6228	78.66%
w/o Sec. 2.5	74.98%	60.89%	0.6228	78.53%

Empirical results demonstrate that MCKD achieves consistent state-of-the-art performance across all ranking and recommendation metrics, delivering significant and comprehensive gains over both conventional multi-task and advanced itinerary-based frameworks, thereby confirming its robust ability to resolve intent ambiguity and signal sparsity in real-world large-scale flight search.

5. System Deployment Design & Online A/B Test





Deployment & Efficiency

- **High Efficiency:** Online inference latency is consistently maintained below 30 ms, meeting stringent real-time responsiveness requirements.
- **Scalable Architecture:** Deployed on a distributed TensorFlow framework, ensuring efficient scalability for large-scale industrial use without sacrificing training efficiency.

Online A/B Test

- **+1.56% CVR Lift :** Achieved a significant +1.56% increase in Conversion Rate (CVR) compared to the production baseline in a two-week live A/B test.
- **+2.14% CTR Lift :** Achieved a significant +2.14% increase in Click Rate (CVR) compared to the production baseline in a two-week live A/B test.
- **Proven Robustness:** Demonstrated superior performance in handling rapid data shifts and sparsity, especially during high-traffic holiday periods.