Contents

Final Selected Features for MAL Prediction Model

Model Performance: $R^2 = 0.7361$ (73.6% variance explained) **Feature Count**: 11 features (reduced from 47) **Selection Criteria**: $p < 0.05$ significance, VIF < 10 (no multicollinearity)

Feature Selection Process

1. **Initial LMM Analysis**: 47 features analyzed with Linear Mixed Model
2. **Statistical Filtering**: 15 features selected with $p < 0.05$
3. **Multicollinearity Removal**: 4 features removed with VIF > 10
4. **Final Set**: 11 features with strong predictive power and no multicollinearity

---

Final 11 Features

1. needs_health_data

- **Type**: Binary (0 or 1)
- **Definition**: Whether the query requires health/fitness data access
- **Examples**:
  - 1: “운동 기록”, “걸음 수”, “칼로리 소모”, “수면 분석”, “건강 데이터를 참고해서”
  - 0: Queries not related to health/fitness data
- **Coefficient**: -0.137 (negative impact on MAL)
- **Rationale**: Health-related analytics are moderately high-stakes; users may accept some latency for correctness but still feel time pressure for daily health decisions.

2. expected_answer_length

- **Type**: Ordinal (0-2)
- **Definition**: Anticipated length/complexity of the answer
  - 0 = Short fact (한 장소/한 시각/횟수)
  - 1 = Medium list (리스트/여러 개의 후보 추천)
  - 2 = Long/complex (요약문, 가계부 작성, 영상 만들기)
- **Examples**:
  - 0: “어디였지?”, “몇 번이나?”

  - 1: “순서대로 알려줘”, “리스트 만들어줘”
  - 2: “요약해줘”, “가계부로 작성해줘”
- **Coefficient**: +0.224 (strong positive impact on MAL)
- **Rationale**: Longer, more elaborate outputs are associated with more computation and higher perceived value, significantly increasing acceptable latency.

3. planning_horizon

- **Type**: Ordinal (0-3)
- **Definition**: Time scope of planning or analysis
  - 0 = Retrospective/analytic (지난달, 작년, 과거 통계/리스트/요약)
  - 1 = Near-future planning (오늘, 내일, 이번 주말)
  - 2 = Medium-term planning (이번 달, 올해 상반기)
  - 3 = Long-term planning (올해 전체, 내년, 장기 습관 개선)
- **Examples**:
  - 0: “지난달 결제 내역 정리”
  - 1: “오늘 점심 메뉴 추천”
  - 2: “이번 달 목표 칼로리”
  - 3: “올해 구독 서비스 추천”
- **Coefficient**: +0.063
- **Rationale**: Retrospective analyses are less time-critical than immediate planning; longer-term planning can tolerate moderate latency if insights are valuable.

4. time_window_length

- **Type**: Ordinal (0-3)
- **Definition**: Approximate span of data to be analyzed
  - 0 = Point in time (어제, 오늘, 특정 날짜)
  - 1 = Days/week (지난 주말, 이번 주)
  - 2 = Weeks/month (지난달, 이번 달)
  - 3 = Months/year+ (작년, 올해, 여름휴가 동안)
- **Examples**:
  - 0: “어제 운동 기록”
  - 1: “이번 주 결제 내역”
  - 2: “지난달 가장 많이 들었던 노래”
  - 3: “작년 이맘때 여행 갔던 곳”
- **Coefficient**: +0.040
- **Rationale**: Longer windows imply more data to scan and aggregate, so users expect and accept higher latency.

5. time_urgency_level

- **Type**: Ordinal (0-2)
- **Definition**: Subjective urgency of the query
  - 0 = Low urgency (지난달/작년, 일반 취향 분석)
  - 1 = Moderate urgency (이번 주/이번 달, 주말 계획)
  - 2 = High urgency (지금, 오늘, 내일, 현재 위치에서)
- **Examples**:

    - 0: “작년에 가장 많이 들었던 음악”
    - 1: “이번 주말 갈만한 맛집”
    - 2: “지금 현재 위치에서 근처 카페”, “오늘 점심 메뉴”
- **Coefficient**: -0.051 (negative impact on MAL)
- **Rationale**: Higher urgency reduces acceptable latency as users want faster responses for immediate decisions.

6. novelty_seeking

- **Type**: Binary (0 or 1)
- **Definition**: Whether the query seeks new/novel recommendations
- **Examples**:
    - 1: “새로운 음악 추천해줘”, “최신 뉴스”, “새로운 콘텐츠”, “트렌드 검색”
    - 0: Queries about known/past items (“내가 들었던”, “갔던 곳”)
- **Coefficient**: -0.165 (negative impact on MAL)
- **Rationale**: Novelty-seeking recommendations are exploratory but users still expect responsive discovery experiences, reducing tolerable latency.

7. requires_aggregation

- **Type**: Binary (0 or 1)
- **Definition**: Whether the query involves counting/aggregating multiple events
- **Examples**:
    - 1: “몇 번이나”, “총 얼마나”, “가장 많이”, “모두 모아서”, “평균”
    - 0: Single-item retrieval queries
- **Coefficient**: +0.042
- **Rationale**: Aggregation implies scanning and combining many records, which users perceive as more complex, slightly increasing acceptable latency.

8. has_comparative_phrase

- **Type**: Binary (0 or 1)
- **Definition**: Whether the query includes comparison language
- **Examples**:
    - 1: “비교해줘”, “평소보다”, “이번 달과 지난달”, “가장 많이”, “더”, “최고”
    - 0: No comparison phrases
- **Coefficient**: +0.275 (strong positive impact on MAL)
- **Rationale**: Comparisons imply multi-period or multi-entity analysis, which users view as significantly more complex, substantially increasing acceptable latency.

9. device_context_implied

- **Type**: Ordinal (0-2)
- **Definition**: Inferred device/usage context
    - 0 = Device-agnostic (general queries, no context clues)
    - 1 = Mobile/on-the-go (현재 위치, 출퇴근 시간, 집 근처, 여기서)
    - 2 = Desktop/home analytics (long retrospective analytics, 가계부 작성)
- **Examples**:

    - 0: “콘텐츠 추천해줘”
    - 1: “현재 위치에서 근처 식당”, “지금 30분 여유”
    - 2: “지난달 카드 결제 내역 정리”, “가계부로 작성해줘”
- **Coefficient**: -0.049 (negative impact on MAL)
- **Rationale**: On-the-go mobile contexts reduce acceptable latency vs. at-home analytics where users can wait longer for richer insights.

10. output_requires_multimedia_creation

- **Type**: Binary (0 or 1)
- **Definition**: Whether the output requires creating multimedia (not just retrieving)
- **Examples**:
    - 1: “영상으로 만들어줘”, “사진 편집해서”, “콜라주 만들어줘”
    - 0: Text or simple retrieval outputs
- **Coefficient**: +0.657 (strongest positive impact on MAL)
- **Rationale**: Multimedia creation is obviously computationally heavy; users expect rendering to take time and thus accept significantly higher latency.

11. social_context_strength

- **Type**: Ordinal (0-2)
- **Definition**: Strength of social context in the query
    - 0 = No social context (personal queries, 나의, 내)
    - 1 = Generic group mention (친구들, 가족, 동료들)
    - 2 = Named individual (성진이, 재희, 나연이, 소연이)
- **Examples**:
    - 0: “내가 자주 들었던 음악”
    - 1: “가족들이랑 갈만한 여행지”
    - 2: “재희랑 갔던 맛집”, “성진이에게 줄 선물 추천”
- **Coefficient**: +0.007 (minimal impact)
- **Rationale**: Stronger social context can increase emotional salience and task complexity; users might tolerate slightly more latency for high-quality personalized results.

Removed Features (VIF > 10)

The following features were removed due to high multicollinearity:

1. **embedding_axis_complexity** (VIF = 62.8)
    - Highly correlated with other complexity indicators
2. **cognitive_load_estimate** (VIF = 33.9)
    - Redundant with combination of other features
3. **task_family** (VIF = 12.5)
    - Information captured by other specific task indicators
4. **requires_personal_history** (VIF = 10.7)
    - Overlaps with domain-specific features

Feature Importance Summary

Top 3 Positive Predictors (increase MAL tolerance): 1. **output_requires_multimedia_creation** (+0.657) - Strongest predictor 2. **has_comparative_phrase** (+0.275) - Complex analysis 3. **expected_answer_length** (+0.224) - Elaborate outputs

Top 3 Negative Predictors (decrease MAL tolerance): 1. **novelty_seeking** (-0.165) - Expect quick discovery 2. **needs_health_data** (-0.137) - Time-sensitive health decisions 3. **time_urgency_level** (-0.051) - Immediate needs

---

Usage Notes

For LLM Feature Extraction

When extracting these features from Korean queries using LLM:

1. **Binary features** (0 or 1): Look for explicit keywords or semantic indicators
2. **Ordinal features**: Assess degree/intensity based on query context
3. **Context clues**:
    - Temporal phrases (지금, 오늘, 지난달)
    - Action verbs (만들어줘, 보여줘, 분석해줘)
    - Comparison words (비교, 가장, 더)
    - Social mentions (names, 친구들, 가족)

Feature Validation

- All features should be extractable from query text alone
- No external context (user history, location) required for extraction
- Features are interpretable and actionable for system design
- Low correlation between features (VIF < 10) ensures stable predictions

---

Model Statistics

- **R²**: 0.7361 (73.6% of MAL variance explained)
- **RMSE**: 1.54 seconds (on log scale: 0.43)
- **ICC**: 0.709 (70.9% of variance from participant individual differences)
- **All features**: Statistically significant ($p < 0.05$)
- **No multicollinearity**: All VIF < 10

References

- Original feature specification: `docs/feature_specification.md`
- VIF analysis: `projects/LMM_model/outputs/vif_analysis.csv`
- Model coefficients: `projects/LMM_model/outputs/lmm_model1_selected_coefficients.csv`
- Selected features list: `projects/LMM_model/outputs/final_selected_features.csv`