RoastFormer: Project Summary for AI Showcase Consideration

Student: Charlee Kraiss **Course**: Generative AI Theory (Fall 2025) **Project**: Transformer-Based Coffee Roast Profile Generation **Repository**: https://github.com/CKraiss18/roastformer

---

Executive Summary

RoastFormer is a **transformer-based generative model** that creates coffee roast profiles (time-series temperature sequences) conditioned on bean characteristics and **desired flavor outcomes**. The novel contribution is **flavor-conditioned generation**—the first model to condition roast profiles on sensory targets (e.g., "berries, floral, citrus"), validated with a **14% performance improvement** over no-flavor baseline.

The project demonstrates successful application of transformer architectures to a **domain-specific sequential generation task** with **small data** (144 profiles), achieving **10.4°F RMSE** on validation. Evaluation revealed **autoregressive exposure bias** (0% physics compliance)—a well-documented challenge—and attempted solutions that failed instructively, providing valuable lessons about post-processing vs training-time fixes.

Showcase Potential: Combines practical domain application (specialty coffee), novel multi-modal conditioning (flavors + bean features), systematic ablation studies, honest evaluation with negative results, and clear course concept integration.

Problem & Motivation

The Real-World Problem

Coffee roasters spend 10-20 experimental roasts (~15 minutes each) per new coffee, working from zero to find an optimal profile. This represents:

- 2-3 hours of experimentation time per coffee
- $200+ in wasted beans and labor per coffee
- Inconsistent results for new roasters

Roasters currently work from experience, simple curve templates, or trial-and-error. No data-driven tools exist for **profile generation** conditioned on sensory outcomes.

The Gap This Fills

Existing work:

- Roast profile databases (static lookup, no generation)
- PID control systems (execute profiles, don't create them)
- Physics-based simulators (complex, require expert tuning)

What's missing: Generative model that learns from real specialty coffee data to create starting profiles conditioned on:

1. Bean characteristics (origin, process, variety, altitude, density)
2. Target roast level (light, medium, dark)
3. **Desired flavor profile** (novel contribution) ← No existing work does this!

Why This Matters (AI Perspective)

This is a **domain-specific sequential generation problem** with interesting constraints:

- **Multi-modal conditioning**: Categorical + continuous + multi-hot flavor features
- **Physics constraints**: Valid roast profiles must respect thermodynamics (monotonicity, bounded heating rates, smooth transitions)
- **Small data regime**: 144 samples from specialty roaster (tests generalization limits)
- **Evaluation challenge**: Standard metrics (RMSE) insufficient; need domain-specific validation (physics compliance)

Broader Impact: Demonstrates transformer applicability beyond NLP/vision to **structured physical processes** with domain constraints.

Technical Architecture

Model Design: Decoder-Only Transformer

Architecture Choice Rationale:

- **Decoder-only** (vs encoder-decoder): Unidirectional causality in roast profiles (temperature at t+1 depends on t, t-1, ...)
- **Autoregressive generation**: Predict next temperature given previous sequence + conditioning
- **Causal masking**: Prevent information leakage from future time steps

Specifications:

```
Model: RoastFormer (Best: d=256)
- Layers: 6 transformer decoder blocks
- Hidden dimension (d_model): 256
- Attention heads: 8
- Feed-forward dimension: 1024 (4x d_model)
- Total parameters: 6,376,673
- Positional encoding: Sinusoidal (ablation tested 3 variants)
- Dropout: 0.1
- Weight decay: 0.01 (critical for small data)
```

Novel Multi-Modal Conditioning Module

Feature Engineering (17 features → unified embedding):

1. **Categorical Features** (5) - Learned embeddings (32-dim each):

- Origin (20 classes: Ethiopia, Colombia, Guatemala, etc.)
- Process (6 classes: Washed, Natural, Honey, Anaerobic, etc.)
- Variety (15 classes: Heirloom, Caturra, Bourbon, etc.)
- Roast Level (4 classes: Expressive Light, Medium, Dark)
- **Flavor Notes** (40 unique) - Multi-hot encoded, projected to 32-dim ← **NOVEL**

2. **Continuous Features** (4) - Normalized, linear projection:

   - Target Finish Temperature (390-430°F)
   - Altitude (1000-2300 MASL)
   - Bean Density Proxy (origin-based)
   - Caffeine Content (variety-based)

3. **Conditioning Mechanism**:

```
categorical_embeds = concat([embed_origin, embed_process, ...,
embed_flavors])
continuous_projected = linear(continuous_features)
condition_vector = concat([categorical_embeds, continuous_projected])

# Cross-attention in each decoder layer
output = self_attention(temp_seq) + cross_attention(temp_seq,
condition_vector)
```

Training Configuration

Optimizer: AdamW (β1=0.9, β2=0.999, weight_decay=0.01) **Learning Rate**: 1e-4 with CosineAnnealingLR (T_max=100) **Loss Function**: MSE (Mean Squared Error) **Batch Size**: 16 **Gradient Clipping**: 1.0 **Early Stopping**: Patience=20 epochs **Regularization**: Dropout (0.1) + weight decay (0.01) + early stopping

Critical Fix: Temperature normalization to [0,1] range

- **Without normalization**: All models collapsed (constant 16°F prediction)
- **With normalization**: 27x faster convergence, all models succeeded
- **Lesson**: Network outputs naturally live near initialization scale (0-10). Raw temps (150-450°F) caused gradient explosion/vanishing.

Data

Source: Scraped Onyx Coffee Lab (2019 US Roaster Champions) - Transparent Coffee Roaster, posts daily roast profiles on website **Size**: 144 roast profiles

- Training: 123 profiles (85%)
- Validation: 21 profiles (15%)

Characteristics:

- Equipment: Loring S70 Peregrine (convection roaster)
- Duration: 7-16 minutes (mean 11.2 min, 1-second resolution)

- Style: Championship-level modern light roasting (72% light roasts)
- Geographic coverage: 20+ coffee origins (Ethiopia 29%, Colombia 19%, etc.)

Novel Contribution: Flavor-Conditioned Generation

The Idea

Hypothesis: Desired flavor outcomes (e.g., "berries", "chocolate", "floral") should guide roast profile generation, as flavor development is the ultimate goal of roasting.

Why This is Novel:

- No existing roast profile generation work conditions on sensory outcomes
- Most work uses only bean metadata (origin, altitude) or target roast level
- Flavors represent the **goal** (what roaster wants to taste), not just **inputs** (what beans are)

Implementation

Flavor Encoding:

- 40 unique flavor notes extracted from Onyx product descriptions
- Multi-hot encoding (profiles have 2-8 flavors each)
- Categories: Fruits (berries, citrus, stone fruit), Florals (jasmine, rose), Chocolate, Nuts/Sugars, Spices
- Projected to 32-dim embedding via learned linear layer

Conditioning:

- Flavor embedding concatenated with other categorical embeddings
- Cross-attention allows model to attend to flavor information at each time step
- Model learns: "For berry + floral flavors, use THIS temperature trajectory"

Validation: Ablation Study

Experiment: Train identical models with/without flavor features

Configuration	Validation RMSE	Improvement
Without flavors	27.2°F	Baseline
With flavors	23.4°F	**+14% better** ✅

Statistical Significance: 3.8°F improvement on 21-sample validation set ($p<0.05$ via paired t-test)

Conclusion: Flavor conditioning provides **measurable performance gain**, validating the novel contribution. Model learns meaningful flavor-profile relationships from data.

Key Results & Findings

Training Success ✅

Model Size Ablation (Surprising Result!):

Model	d_model	Params	Val RMSE	Params/Sample
Small	d=32	202,945	43.8°F	1,650:1
Medium-S	d=64	605,633	23.4°F	4,925:1
Medium	d=128	2,044,545	16.5°F	16,625:1
Large	**d=256**	**6,376,673**	**10.4°F** ✅	**51,843:1**

Surprising Finding: Largest model (d=256) with 6.4M parameters achieved **best performance** despite 51,843:1 parameter-to-sample ratio.

Initial Hypothesis: "d=256 will overfit on 123 samples" ❌ **Reality**: d=256 performed best ✅ **Why I Was Wrong**: Normalization was the fundamental bug. With proper regularization (dropout, weight decay, early stopping), larger models leverage capacity to learn complex roast dynamics without overfitting.

Lesson: Being experimentally wrong taught more than being theoretically correct.

Positional Encoding Ablation

Experiment: Compared 3 positional encoding methods

Method	Val RMSE	Notes
Sinusoidal	**23.4°F** ✅	Classic Vaswani et al. 2017
RoPE	28.1°F	From my RoPE presentation!
Learned	43.8°F	Overfits on small data

Interesting: RoPE (more complex) performed worse than sinusoidal (simpler) on small data. **Lesson**: Classic methods win in low-data regimes. Complexity ≠ better performance when data-limited.

Flavor Conditioning Validation (Novel Contribution)

- **With flavors**: 23.4°F RMSE
- **Without flavors**: 27.2°F RMSE
- **Improvement**: 3.8°F (14% better) ✅

Novel contribution validated with statistical significance.

Evaluation Challenge: Autoregressive Exposure Bias

Generation Performance

Metrics (Unconstrained generation on 10 validation samples):

Metric	Value	Assessment
MAE	25.3°F	Reasonable temp accuracy

Metric	Value	Assessment
RMSE	29.8°F	3x worse than training (10.4°F)
Finish Temp Accuracy (±10°F)	50%	Decent
Physics Compliance	0%	❌ Problem!

Physics Compliance Breakdown:

- Monotonicity (post-turning point): 0.0% ❌
- Bounded RoR (20-100°F/min): 28.8% ⚠
- Smooth Transitions (<10°F/s): 98.7% ✅
- Overall Valid: 0.0% ❌

Root Cause: Exposure Bias

The Problem:

- **During training**: Model sees real previous temperatures (teacher forcing) → learns patterns ✅
- **During generation**: Model sees own predictions → errors compound → physics violations ❌

Training vs Generation Gap:

- Training RMSE: 10.4°F (with teacher forcing)
- Generation MAE: 25.3°F (autoregressive)
- Gap: 2.4x worse

This is the **autoregressive exposure bias problem** (Bengio et al., 2015) - well-documented in sequence generation literature.

Attempted Solution: Physics-Constrained Generation (LESSONS LEARNED)

Hypothesis

"Enforcing physics constraints during generation (monotonicity, bounded heating rates) should improve compliance while maintaining accuracy."

Implementation

Constraints Applied:

1. Monotonic increase after turning point (no cooling)
2. Bounded heating rates (20-100°F/min)
3. Smooth transitions (<10°F/s)
4. Physical temperature bounds (250-450°F)

Results: FAILED ❌

Metric	Unconstrained	Constrained	Change

Metric	Unconstrained	Constrained	Change
MAE	25.3°F	113.6°F	**+88.3°F (4.5x worse)** ❌
Finish Temp MAE	13.95°F	86.67°F	**+72.7°F worse** ❌
Monotonicity	0.0%	100.0%	+100% ✅
Bounded RoR	28.8%	0.0%	**-28.8% (worse!)** ❌

Visual Evidence: Constrained generation produced linear ramps (330°F → 500°F straight lines) instead of realistic curves.

Why It Failed: Root Cause Analysis

The Fundamental Issue: Constraints fight against the model's learned behavior.

What the Model Learned (during training with teacher forcing):

- Temperature patterns that include non-monotonic segments
- Heating rates occasionally outside 20-100°F/min bounds
- Complex curve dynamics (drying dip, maillard acceleration, development slowdown)

What Constraints Force (during generation):

- Strictly monotonic increases → eliminates learned curve features
- Hard bounds on RoR → model tries to predict natural dynamics, constraints override
- Result: Model and constraints in conflict → linear ramps, not curves

The Lesson: **Post-processing constraints cannot fix training issues.**

Solutions must address the root cause (training process), not symptoms (generation output):

- ✅ **Scheduled Sampling** (Bengio et al., 2015): Train with model's own predictions, not just teacher forcing
- ✅ **Physics-Informed Loss Functions**: Add penalty terms for violations during training
- ✅ **Non-Autoregressive Generation**: Diffusion models (no error accumulation)

Value of This "Failure"

This negative result demonstrates:

1. **Scientific maturity**: Documented failed approach honestly
2. **Root cause understanding**: Identified why post-processing fails
3. **Literature grounding**: Connected to proper solutions (scheduled sampling)
4. **Critical thinking**: Post-hoc fixes ≠ training-time solutions

For AI showcase: This is more valuable than claiming everything worked. Shows real research process.

Learning Journey: Debugging Story

Initial Failure: Model Collapse

Problem: ALL 10 initial models predicted constant 16°F (total failure)

Systematic Debugging Process:

1. Tried smaller models (d=32, d=64) → still failed
2. Reduced learning rate (1e-5) → still failed
3. Analyzed training logs → found gradient explosion
4. Examined data distributions → discovered scale mismatch

Critical Bug #1: Missing Normalization

Root Cause: Temperature scale mismatch

- **Targets**: 150-450°F (raw temperatures)
- **Network outputs**: 0-10 (typical initialization scale)
- **Result**: Gradient explosion/vanishing, learning impossible

Fix: Normalize temperatures to [0, 1] range

```
temp_normalized = (temp - temp.min()) / (temp.max() - temp.min())
```

Impact: 27x faster convergence, all models succeeded

Lesson: Neural networks naturally output values near initialization scale. Asking for raw temps (150-450°F) breaks gradient flow. Normalization is fundamental, not optional.

Critical Bug #2: Wrong Hypothesis About Capacity

Initial Belief: "6.4M parameters will overfit on 123 samples"

Experiments: Trained d=32, d=64, d=128, d=256 with proper regularization

Result: d=256 achieved **best performance** (10.4°F RMSE)

Why I Was Wrong:

- Normalization was THE critical bug
- With proper regularization (dropout, weight decay, early stopping), capacity helps
- Larger models learn complex roast dynamics better
- 51,843:1 ratio is fine with modern regularization techniques

Lesson: Empirical validation > theoretical assumptions. The experiment proved my hypothesis wrong—and that's valuable learning.

Course Integration (Generative AI Theory)

Week 2: Neural Network Fundamentals

Applied: Temperature normalization (critical bug fix) **Lesson**: "Networks output values near initialization scale. Normalization isn't a trick—it's fundamental to gradient flow."

Week 4: Autoregressive Modeling & Exposure Bias

Applied: Sequential temperature generation with teacher forcing **Challenge**: Exposure bias identified through evaluation **Lesson**: "Training with real sequences doesn't prepare model for generating from own predictions. Literature-backed solutions: scheduled sampling."

Week 5: Transformer Architecture & Positional Encodings

Applied: Compared sinusoidal, RoPE, learned positional encodings **Result**: Sinusoidal > RoPE > Learned (opposite of complexity order) **Lesson**: "Classic methods win in small-data regimes. Presented RoPE in class, then validated that simpler beats complex with limited data."

Week 6-7: Conditional Generation (Multi-Modal Features)

Applied: Flavor-conditioned generation (novel contribution) **Result**: 14% improvement validates approach **Lesson**: "Task-relevant conditioning (flavors) improves generation quality. Multi-modal features (categorical + continuous + multi-hot) require careful encoding."

Week 8: Small-Data Regime Strategies

Applied: Heavy regularization (dropout, weight decay, early stopping) **Surprising Result**: d=256 with 51,843:1 ratio achieved best performance **Lesson**: "Normalization + regularization > capacity limits. Being wrong experimentally taught more than being right theoretically."

Week 9: Evaluation Methodology & Domain-Specific Metrics

Applied: Physics-based validation (monotonicity, bounded RoR, smoothness) **Finding**: Standard metrics (RMSE) don't capture domain constraints **Lesson**: "Generic metrics mislead. Needed domain-specific validation to reveal exposure bias problem. Honest reporting of limitations > hiding failures."

Limitations & Future Work

Current Limitations (Honest Assessment)

1. **Exposure Bias** (Critical):
    - 0% physics compliance during generation
    - Profiles violate roasting physics (non-monotonic, unbounded RoR)
    - **NOT production-ready** - requires human validation

2. **Single-Roaster Bias** (Critical):
    - All 144 profiles from Onyx Coffee Lab only
    - Model learns "Onyx's championship style" not "how to roast"
    - Equipment-specific (Loring S70 convection only)
    - **Key insight**: Even 500 Onyx profiles wouldn't fix this—need 10+ diverse roasters

3. **Light Roast Bias**:
    - 72% light roasts, only 2% dark
    - May generate poor dark roast profiles
4. **Small Dataset**:
    - 144 samples limits pattern diversity
    - Amplifies exposure bias problem

Proper Solutions (Literature-Backed)

1. **Scheduled Sampling** (Bengio et al., 2015)
    - Gradually transition from teacher forcing to model predictions during training
    - Addresses exposure bias at the source
    - **Expected impact**: 25→15°F MAE, 0→80%+ physics compliance
2. **Physics-Informed Loss Functions**
    - Add penalty terms for physics violations to training loss
    - Model learns to respect constraints
    - Example: `loss = mse_loss + λ₁*monotonicity_penalty + λ₂*ror_penalty`
3. **Multi-Roaster Dataset** (Most Critical!)
    - 500+ profiles from **10+ diverse roasters** (not 500 from Onyx!)
    - Equipment diversity: Loring, Probat, Diedrich (drum), Sivetz (fluid bed)
    - Style diversity: Nordic light, traditional medium, French dark, espresso
    - Geographic diversity: US, Europe, Asia, Africa roasting cultures
    - **Key lesson**: Diversity > scale. 200 from 10 roasters > 500 from one roaster
4. **Non-Autoregressive Architectures**
    - Diffusion models for roast profile generation
    - Generate entire sequence at once (no error accumulation)
    - Eliminates exposure bias entirely
5. **Duration Prediction Module**
    - Current: User specifies duration (design choice, like target temp)
    - Future: Model predicts optimal duration for coffee
    - "This dense Ethiopian at 2100m needs 11.5 min for light roast"

Why This is Showcase-Worthy

1. Novel Domain Application ✅

- Transformers applied to **domain-specific physical process** (roasting)
- Beyond NLP/vision → structured time-series with physics constraints
- Demonstrates generative AI for **practical specialty domain** (coffee)

2. Novel Technical Contribution ✅

- **Flavor-conditioned generation** (first in roast profiling)
- Multi-modal conditioning (categorical + continuous + multi-hot)
- Validated with **14% improvement** (statistically significant)

3. Small-Data Success ✅

- 144 samples, 6.4M parameters (51,843:1 ratio) → 10.4°F RMSE
- Demonstrates **proper regularization** overcomes data scarcity
- Surprising result: Larger model won (opposite of prediction)

4. Honest Scientific Process ✅

- **Documented failures**: Constrained generation attempt (MAE 4.5x worse)
- **Root cause analysis**: Why post-processing fails
- **Literature grounding**: Proper solutions cited (scheduled sampling)
- Shows **research maturity** > claiming everything worked

5. Systematic Ablation Studies ✅

- Model size: d=32, 64, 128, 256 (comprehensive)
- Positional encodings: Sinusoidal, RoPE, Learned (theory → practice)
- Flavor conditioning: With/without (validates contribution)
- Demonstrates **experimental rigor**

6. Strong Course Integration ✅

- Applied concepts from **6 weeks** (Week 2, 4, 5, 6-7, 8, 9)
- Each experiment tied to theoretical concept
- Shows **depth of understanding** beyond implementation

7. Clear Future Work ✅

- Identified specific problems (exposure bias, roaster diversity)
- Literature-backed solutions (scheduled sampling, physics-informed losses)
- Actionable next steps (multi-roaster dataset, diffusion models)

8. Practical Potential 🚀

- Addresses real problem (10-20 experimental roasts per coffee)
- Clear user value ($200+ saved, 2-3 hours reduced)
- Path to production (with proper solutions applied)

Technical Highlights for AI Audience

Architecture Choices:

- Decoder-only (causal structure matches problem)
- Cross-attention for multi-modal conditioning (17 features → unified embedding)

- Sinusoidal PE > RoPE on small data (validated empirically)

Training Innovations:

- Temperature normalization critical (27x speedup)
- Heavy regularization enables large models on small data
- Physics-aware evaluation (domain metrics > generic metrics)

Novel Conditioning:

- Flavor features (multi-hot, 40 classes) as generation targets
- 14% improvement validates approach
- Opens research direction: sensory outcome conditioning

Evaluation Rigor:

- Standard metrics (RMSE: 10.4°F training, 25.3°F generation)
- Domain metrics (physics compliance: 0% - exposure bias)
- Negative results documented (constrained generation failure)

Research Contributions:

1. Flavor-conditioned roast profile generation (novel)
2. Transformer application to physics-constrained sequential generation
3. Small-data regime validation (6.4M params, 123 samples)
4. Documented exposure bias in domain application with attempted solutions

---

Repository & Documentation

GitHub: https://github.com/CKraiss18/roastformer

Comprehensive Documentation (35+ files):

- `docs/MODEL_CARD.md` - Complete model documentation
- `docs/DATA_CARD.md` - Dataset documentation & ethics
- `docs/EVALUATION_FINDINGS.md` - Complete evaluation + lessons learned
- `docs/COMPREHENSIVE_RESULTS.md` - All ablation studies
- `docs/METHODOLOGY_COURSE_CONNECTIONS.md` - Course concept mapping
- `docs/RUBRIC_COURSE_MAPPING.md` - Rubric alignment (115/125 pts projected)

Code:

- `train_transformer.py` - Complete training pipeline
- `evaluate_transformer.py` - Evaluation suite
- `generate_profiles.py` - Profile generation from features
- `src/model/transformer_adapter.py` - Model architecture

Results Package:

- All training experiment results (7 ablations)
- Evaluation metrics & visualizations

- Real vs generated profile comparisons
- Physics compliance analysis

Final Thoughts

RoastFormer demonstrates that **transformers can learn domain-specific physical processes** from real specialty data, with **proper conditioning on sensory outcomes** (novel contribution). The project showcases **systematic experimental methodology** (7 ablations), **honest scientific reporting** (documented failures), and **strong course integration** (6 weeks of concepts).

The current limitations (0% physics compliance, single-roaster bias) are **honestly documented with literature-backed solutions**. The surprising results (d=256 won, normalization critical, constrained generation failed) provide **valuable learning** beyond a "perfect" project.

For an AI showcase, this offers:

- **Novel domain**: Coffee roasting (practical, relatable, interesting)
- **Technical depth**: Multi-modal conditioning, small-data success, physics constraints
- **Research maturity**: Systematic ablations, negative results, root cause analysis
- **Story arc**: Failure → debugging → success → new challenges → lessons learned

I believe this would engage an AI audience by demonstrating **transformers beyond NLP/vision**, **small-data techniques**, and **honest research process**.

Thank you for considering RoastFormer for the showcase. I'd greatly appreciate your feedback!
