Slide 1 — Project Review

Motivation & problem

- Clinical notes frequently contain follow-up actions (tests, imaging, consults) with relative timing (e.g., “within 2 weeks”).
- Goal: automatically extract scheduled actions + timing and convert to a concrete due date so it can be added to a calendar.
- Two approaches: (A) a deterministic NLP pipeline (BioBERT + date normalization) and (B) a generative LLM (Llama 3) that outputs actions + due dates directly.

Project specification (IO)

Input

Free-text clinical note (includes visit date + plan section)

Output

List of { action, period_date }

What changed vs proposal

- Expanded from “extract action + number of days” → end-to-end “action + computed due date”.
- Added a direct Llama 3 fine-tuning baseline to compare against the deterministic pipeline.
- Upgraded synthetic data generation to include deterministic period_date labels and character-level spans for BIO training.

Contributions

- A 10k-note synthetic benchmark with realistic variation across specialties, styles, and timing phrases, plus span labels.
- Multi-task BioBERT: multi-label action detection + BIO time-span tagging, followed by deterministic date normalization.
- A controlled comparison: pipeline vs generative model that must output the final period dates directly.

Slide 2 — Research Papers (Literature)

Title / Year	Task	Methods	Data	Results	Relation to Project
"BioBERT: a pre-trained biomedical language representation model" (Lee et al., 2020)	Biomedical NER & Relation Extraction.	Domain-specific pre-training of BERT on PubMed/PMC.	PubMed, PMC, **NCBI Disease**.	**F1: 80-90%** (SoTA on biomedical NER).	Provides the foundational backbone (BioBERT v1.1) for our token classification, proving domain-specific encoders outperform generic BERT.
"Clinical Temporal Relation Extraction with Probabilistic Soft Logic" (Zhou et al., 2021)	Extracting temporal links between events.	Hybrid: Neural Network + Probabilistic Soft Logic (PSL).	**i2b2-2012** (Standard clinical corpus).	**F1: 0.68** (SoTA for relations).	Validates our **Neuro-Symbolic architecture**: combining neural extraction (BioBERT) with symbolic logic (Python rules) to enforce timeline consistency.
"Relation Extraction in Biomedical Texts Based on Multi-Head Attention" (Li et al., 2022)	Extracting complex clinical entities & relations.	**Bi-LSTM** with **Multi-Head Attention** mechanisms.	Biomedical Literature / Clinical Text.	**Precision: 84.5%** on relation tasks.	Justifies our Phase 4 plan to "train attention weights." It proves that attention mechanisms handle long-distance dependencies (e.g., Action... [20 words] ... Date) better than standard RNNs.

Slide 3 — Dataset (10k synthetic outpatient notes)

What we generated & labeled

- 10,000 notes across 5 specialties (Orthopedic, Cardio/Pulm, GI, Neuro, General).
- Each note contains a visit date and optionally a plan header; 0–2 scheduled items per note.
- Ground truth per scheduled item: action (closed set), period_text, computed period_date, plus character spans (action/time) for BIO training.

Generation technique

- Skeleton-first: sample specialty/topic/styles, visit date, actions, and deterministic timing phrase templates.
- LLM writes the narrative note; plan section uses a random header and includes exactly the allowed scheduled items.
- Deterministic period_date label computed with dateparser using RELATIVE_BASE = visit date .

EDA snapshot

Samples

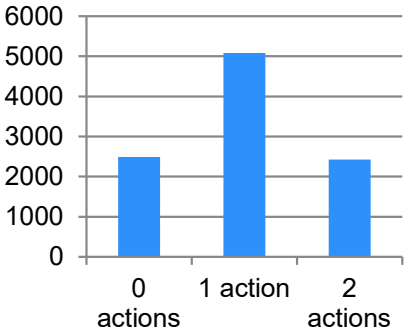
10k

balanced specialties

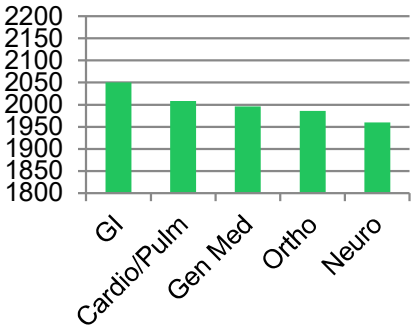
Mean length

141

words / note



num_actions distribution



specialty distribution

Slide 3 — Dataset (10k synthetic outpatient notes)

Generating diverse dataset

```
SPECIALTY_MAPPING = {
    "Orthopedic": [
        "MRI", "Physical Therapy", "CT Scan", "X-Ray", "Orthopedic Consult", "Joint Injection"
    ],
    "Cardiovascular / Pulmonary": [
        "Echocardiogram", "Stress Test", "Holter Monitor", "Pulmonary Function Test",
        "Cardiac MRI", "Cardiology Consult"
    ],
    "Gastroenterology": [
        "Endoscopy", "Colonoscopy", "Stool Antigen Test", "Abdominal Ultrasound",
        "GI Consult", "Breath Test"
    ],
    "Neurology": [
        "CT Scan", "EEG", "MRI Brain", "EMG", "Neurology Consult", "Sleep Study"
    ],
    "General Medicine": [
        "Blood Test", "X-Ray", "Vaccination", "Annual Physical", "Urinalysis", "Lipid Panel"
    ]
}
```

```
TOPICS = {
    "Orthopedic": [
        "ACL Tear", "Rotator Cuff Injury", "Lumbar Herniated Disc", "Ankle Sprain",
        "Carpal Tunnel Syndrome", "Hip Osteoarthritis", "Meniscus Tear", "Tennis Elbow",
        "Plantar Fasciitis", "Distal Radius Fracture",
        "Sciatica", "Scoliosis", "Bunion (Hallux Valgus)", "Patellar Tendonitis",
        "Shoulder Dislocation"
    ],
    "Cardiovascular / Pulmonary": [
        "Atrial Fibrillation", "COPD Exacerbation", "Acute Bronchitis", "Hypertension",
        "Mitral Valve Prolapse", "Pneumonia", "Congestive Heart Failure", "Deep Vein Thrombosis",
        "Asthma Attack", "Pericarditis",
        "Pulmonary Embolism", "Coronary Artery Disease", "Aortic Stenosis", "Bradycardia",
        "Syncope/Fainting"
    ],
    "Gastroenterology": [
        "GERD", "Irritable Bowel Syndrome", "Crohn's Disease", "Ulcerative Colitis",
        "Gallstones", "Celiac Disease", "Peptic Ulcer", "Diverticulitis",
        "Hemorrhoids", "Liver Cirrhosis",
        "Barrett's Esophagus", "Acute Gastritis", "Hiatal Hernia", "Chronic Pancreatitis",
        "Hepatitis C"
    ],
    "Neurology": [
        "Migraine with Aura", "Epilepsy/Seizure", "Multiple Sclerosis", "Parkinson's Disease",
        "Ischemic Stroke", "Carpal Tunnel (Neuro view)", "Vertigo/BPPV", "Alzheimer's/Dementia",
        "Bell's Palsy", "Neuropathy",
        "Myasthenia Gravis", "Trigeminal Neuralgia", "Huntington's Disease", "Guillain-Barre Syndrome",
        "Restless Leg Syndrome"
    ],
    "General Medicine": [
        "Type 2 Diabetes", "Seasonal Influenza", "Hypothyroidism", "Urinary Tract Infection",
        "Anemia", "Vitamin D Deficiency", "Hypertension", "Annual Physical",
        "Lyme Disease", "Gout",
        "Hyperlipidemia (High Cholesterol)", "Fibromyalgia", "Chronic Fatigue Syndrome", "Infectious Mononucleosis",
        "Osteoporosis"
    ]
}
```

Slide 3 — Dataset (10k synthetic outpatient notes)

Diverse note style

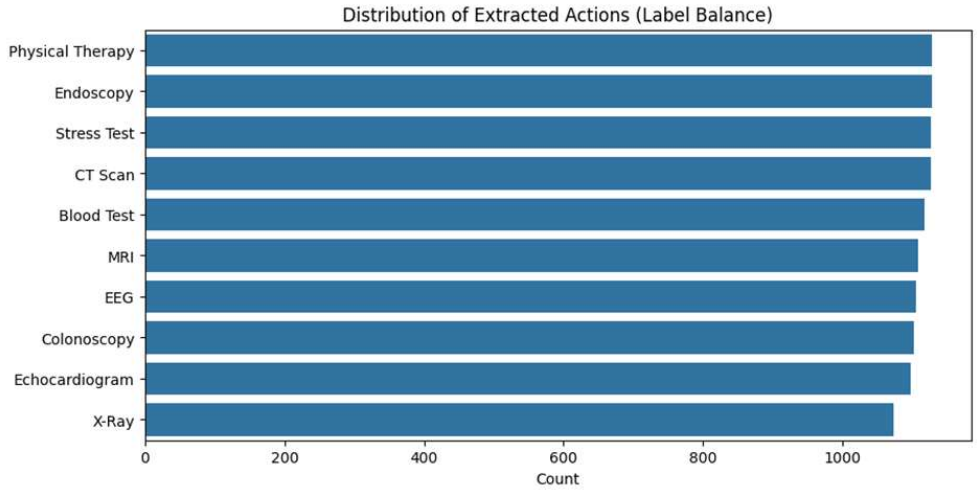
```
STYLE_FEATURES_DICT = {
    "Narrative Voice": [
        "First Person ('I examined the patient', 'I recommend')",
        "Third Person ('Patient presents with', 'It is recommended')"
    ],
    "Temporal Precision": [
        "Vague ('History of surgery years ago', 'pain for a while')",
        "Precise ('Surgery on 12/05/2020', 'pain started at 2 PM')"
    ],
    "Certainty": [
        "Definitive ('Patient has pneumonia')",
        "Hedging/Uncertain ('Findings suggestive of possible pneumonia', 'cannot rule out')"
    ],
    "Formality": [
        "Formal/Academic (Complete sentences, proper grammar)",
        "Casual/Direct (Conversational, simple sentence structures)",
        "Standard Clinical (Professional but concise)"
    ],
    "Detail Level": [
        "Highly Detailed/Verbose (Explains rationale, describes scene)",
        "Abbreviated/Telegraphic (Notes style, fragments)",
        "Standard (Balanced)"
    ],
    "Tone": [
        "Polite/Empathetic ('Patient is a pleasant 45yo...')",
        "Clinical/Detached (Just the facts)",
        "Direct/Urgent ('Patient in distress, immediate action required')"
    ],
    "Terminology": [
        "Simple Language (Patient-friendly terms like 'heart attack')",
        "Heavy Medical Jargon (terms like 'myocardial infarction')",
        "Mixed (Standard EHR style)"
    ],
    "Structure": [
        "Standard SOAP Headers (Subjective: ... Objective: ...)",
        "Minimal Headers (HPI: ... PE: ... Imp: ...)",
        "Run-on Paragraph (No clear section breaks, one block of text)",
        "Letter Format ('Dear Dr. X, thank you for referring...')",
        "Bullet Points (Heavy use of lists for symptoms/plan)"
    ],
    "Imperfections": [
        "Perfect Grammar (textbook quality)",
        "Slight Shorthand (Standard abbreviations like 'pt', 'yo', 'hx')",
        "Heavy Shorthand (Aggressive abbrv: 'c/o CP, SOB, N/V, rec MRI')",
        "Dictation Style (Occasional missing punctuation, run-on sentences)",
        "Minor Typos (Simulated fast typing errors like 'pateint' or 'sweling')"
    ],
    "Clinician Persona": [
        "Defensive Medicine (Over-explaining rationale to justify decisions)",
        "Action-Oriented (Brief history, very detailed plan)",
        "Burnout/Hasty (Minimum viable documentation, very brief)",
        "Educator (Explaining the 'why' behind the diagnosis)"
    ]
```

Slide 3 — Dataset (10k synthetic outpatient notes)

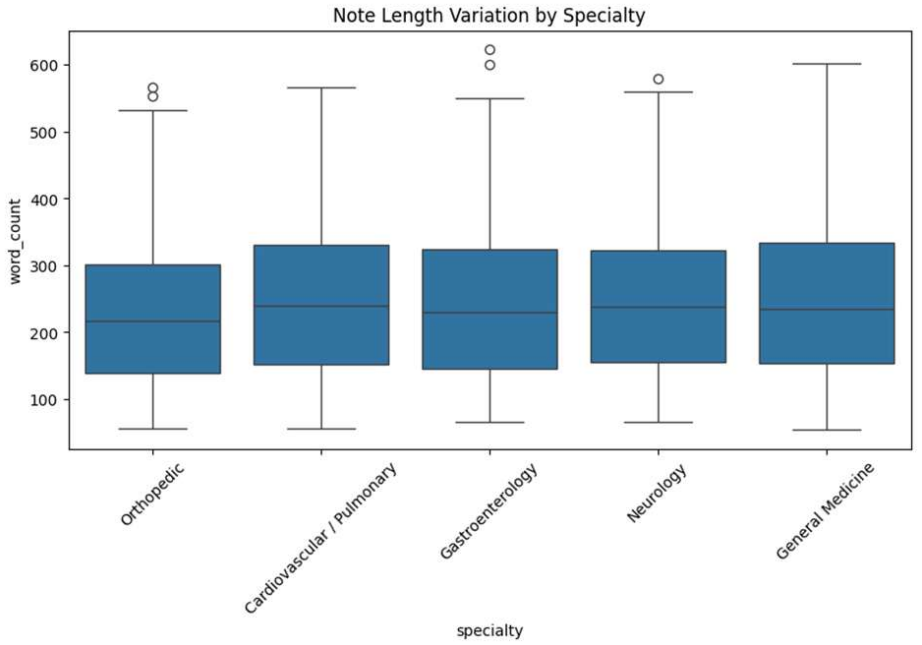
Style Prompt Affect

```
--- Validation: Does the Style Prompt affect Word Count? ---
style_category
Detailed       357.652609
Standard       236.156204
Telegraphic    135.943165
Name: word_count, dtype: float64
```

Actions Distribution

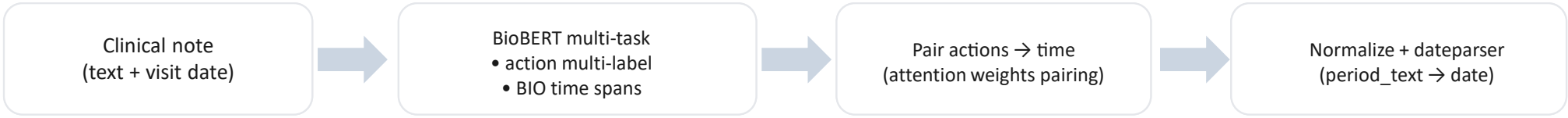


Note Length Per Specialty



Slide 4 — Baseline solution & results

Baseline pipeline



Comparison baseline: fine-tune Llama 3 to output JSON with {action, period_date} directly (no explicit dateparser step).

BioBERT training results

[Debug] Truth: [0. 0. 1. 0. 0. 0. 0. 0. 0. 0.] | Pred: [0. 0. 1. 0. 0. 0. 0. 0. 0. 0.]

Epoch 10: Action F1: 0.7893 | NER Accuracy: 0.9997

- .

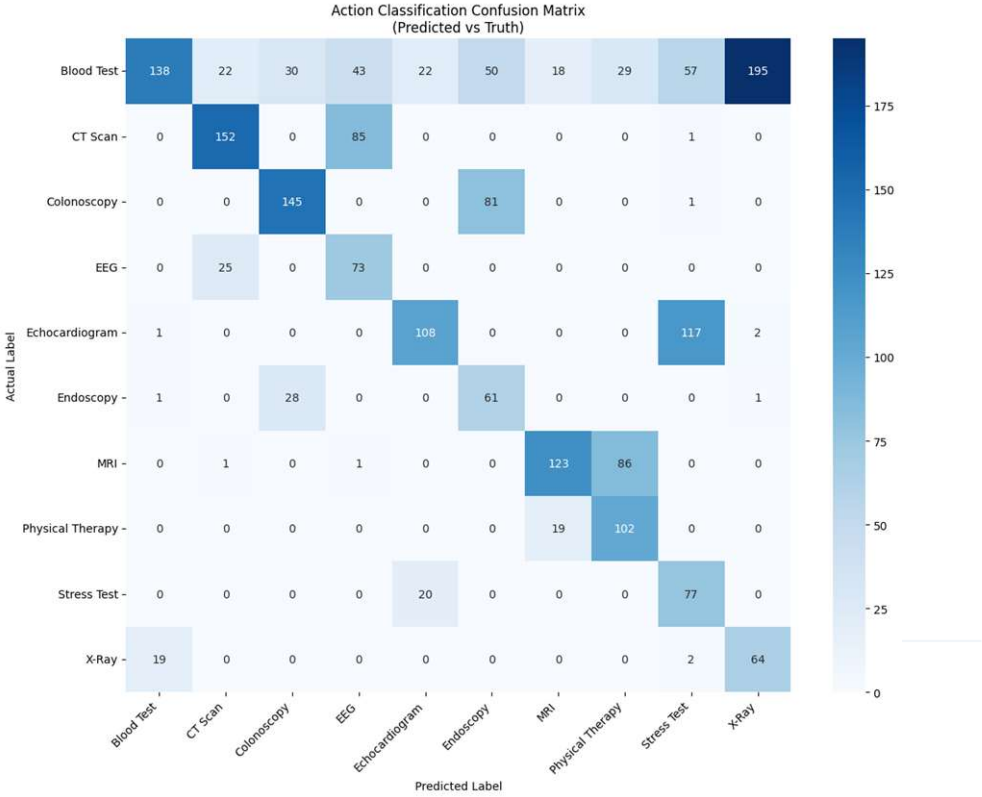
Evaluation Results: BioBERT Performance

Training Metrics
Final Epoch (10): The model achieved strong convergence with an **Action F1 Score of 0.7893** and a near-perfect **NER Accuracy of 0.9997**.
Sample Prediction: Truth: [0, 0, 1, ...] vs Pred: [0, 0, 1, ...]. The alignment of vectors confirms the model successfully learned to encode the action classes.
Threshold Optimization & Robustness We analyzed the decision boundary to maximize F1 performance.
Stability Analysis: The model demonstrated remarkable robustness. Performance remained stable (F1 ~0.79) across a wide threshold range (0.20 – 0.60).
Interpretation: This wide "safe zone" indicates the model is confident in its predictions. It is not "guessing" with low probabilities; rather, it creates a distinct decision boundary between correct and incorrect classes.

Slide 4 — Baseline solution & results

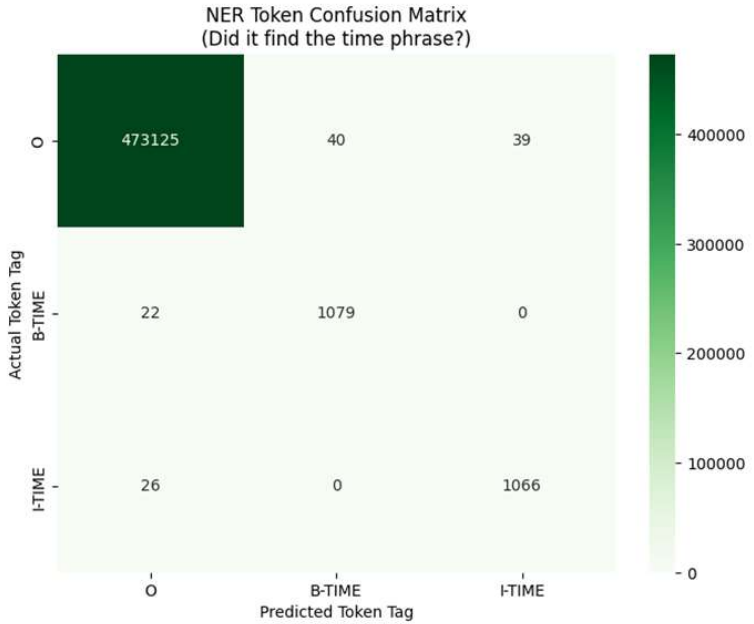
BioBERT training results

successes: The diagonal dominance in the confusion matrix shows high accuracy for distinct procedures like **CT Scans (152 correct)** and **Colonoscopy (145 correct)**.
Confusion Clusters: Errors are not random. The model struggles with semantically similar pairs, such as **Echocardiogram vs. Stress Test**. This suggests the model understands the *domain* (Cardiology) but occasionally misinterprets the specific *modality*.



Error Analysis: Inspecting the Confusion

Time Span Extraction (NER)
Precision: The NER Token Confusion Matrix reveals an extremely low false-positive rate.
Interpretation: The model almost never "hallucinates" time where there is none (O -> B-TIME errors are rare). It effectively ignores the vast majority of non-temporal text (473,125 'O' tokens correctly identified), proving it has learned to focus strictly on temporal signals.



Slide 5 — Plan (scope → milestones → expected outcomes)

What remains (technical roadmap)

- Harden the dataset to better reflect deployment: decoy times in history, past tests vs future scheduling, and occasional paraphrases.
- Train and evaluate Llama 3 on the same data split (predict actions + period_dates directly) to compare robustness and failure modes.
- Stress tests: out-of-template time phrases, longer notes, multiple plan sections, and “problem-oriented” notes where plan is not strictly at the end.

Milestones (table)

Week	phase	Technical Scope & Models	Expected Outcome
W1	Advanced Data: Increase realism & complexity	Inject "history" distractors, past dates, and multi-action plans.	dataset with adversarial examples.
W2	Model Enhancement: Improve logic & linking capabilities	**Attention Mechanism Tuning** & **Llama 3 Fine-tuning**.	Robust handling of multi-action pairing.
W3	Final Evaluation: Comparative analysis of architectures.	**BioBERT vs. Llama 3 vs. GPT-4o (Zero-Shot)**.	Final report on "Discriminative vs. Generative" performance.
W4	Presentation: Synthesize findings and methodology.	Slide deck, code repository.	Final Presentation