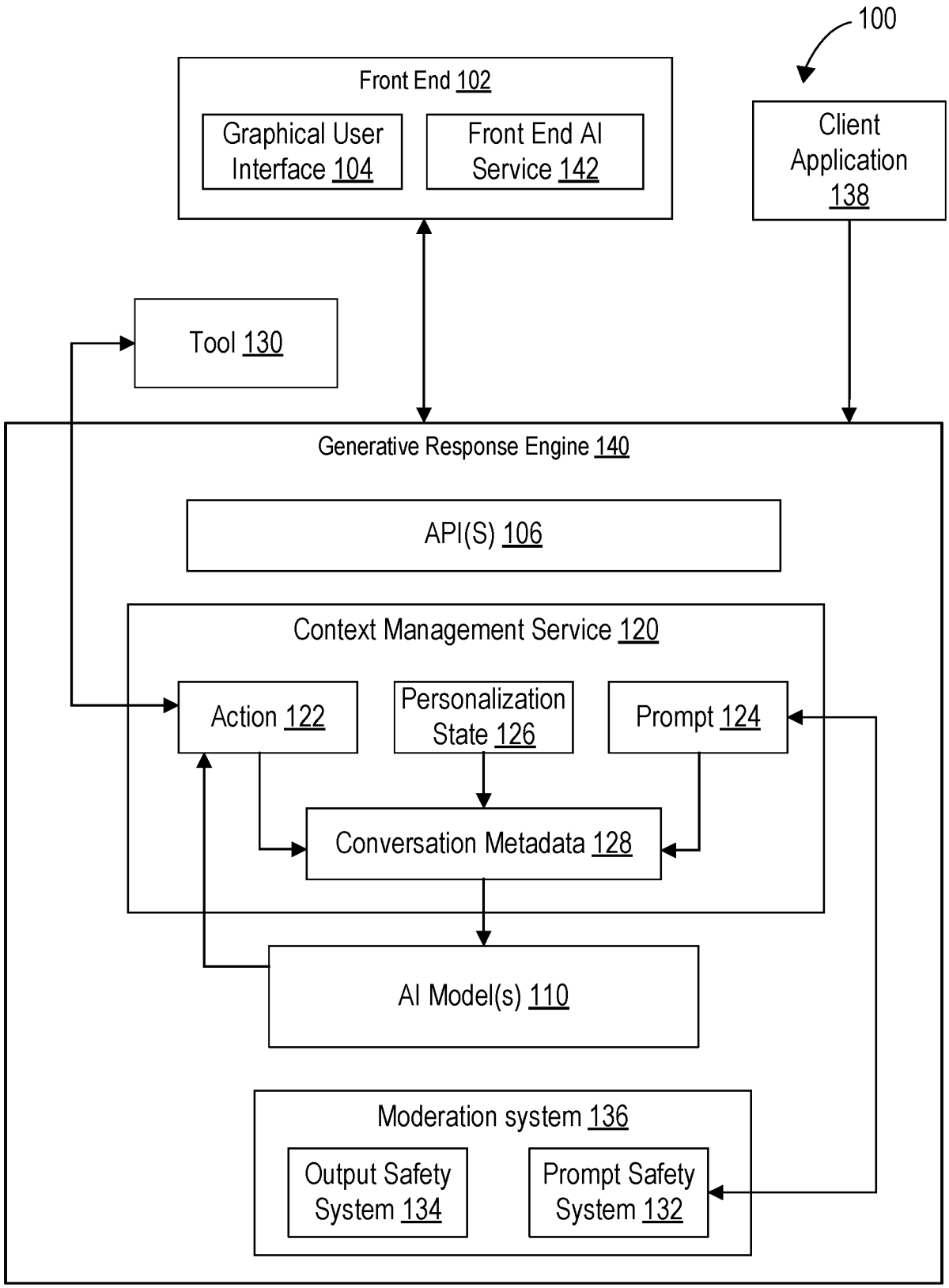

100
Front End 102
Graphical User Interface 104
Front End AI Service 142
Client Application 138
Tool 130
Generative Response Engine 140
API(S) 106
Context Management Service 120
Action 122
Personalization State 126
Prompt 124
Conversation Metadata 128
AI Model(s) 110
Moderation system 136
Output Safety System 134
Prompt Safety System 132

FIG. 1

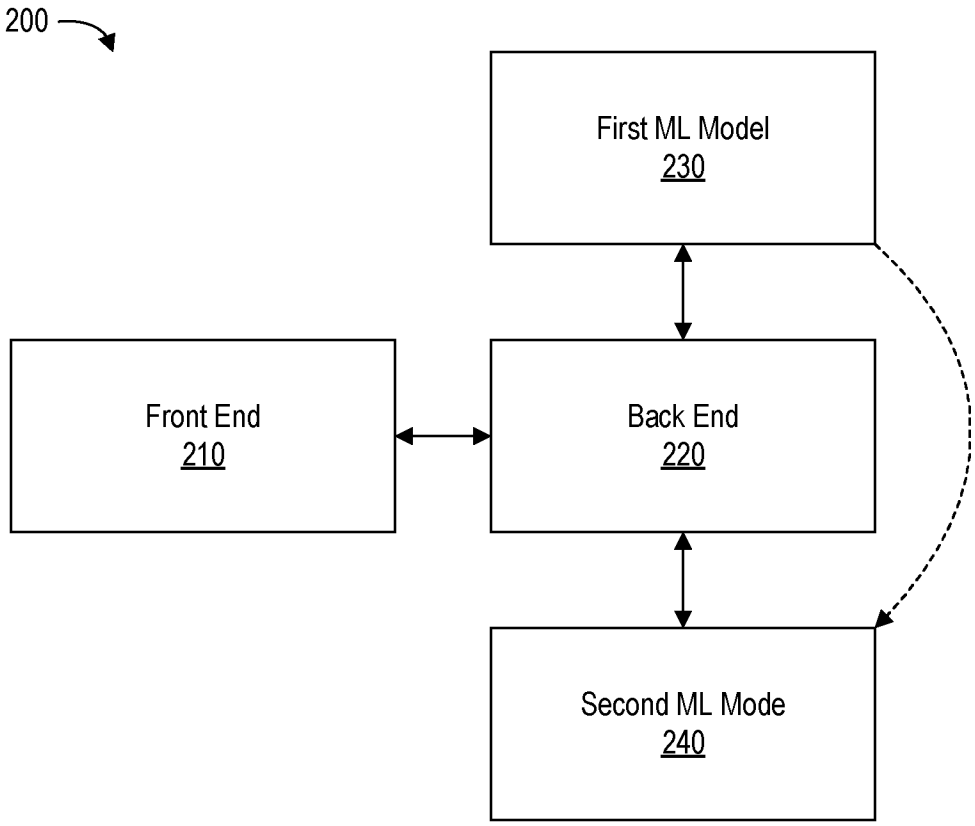

200
First ML Model
230
Front End
210
Back End
220
Second ML Mode
240

FIG. 2

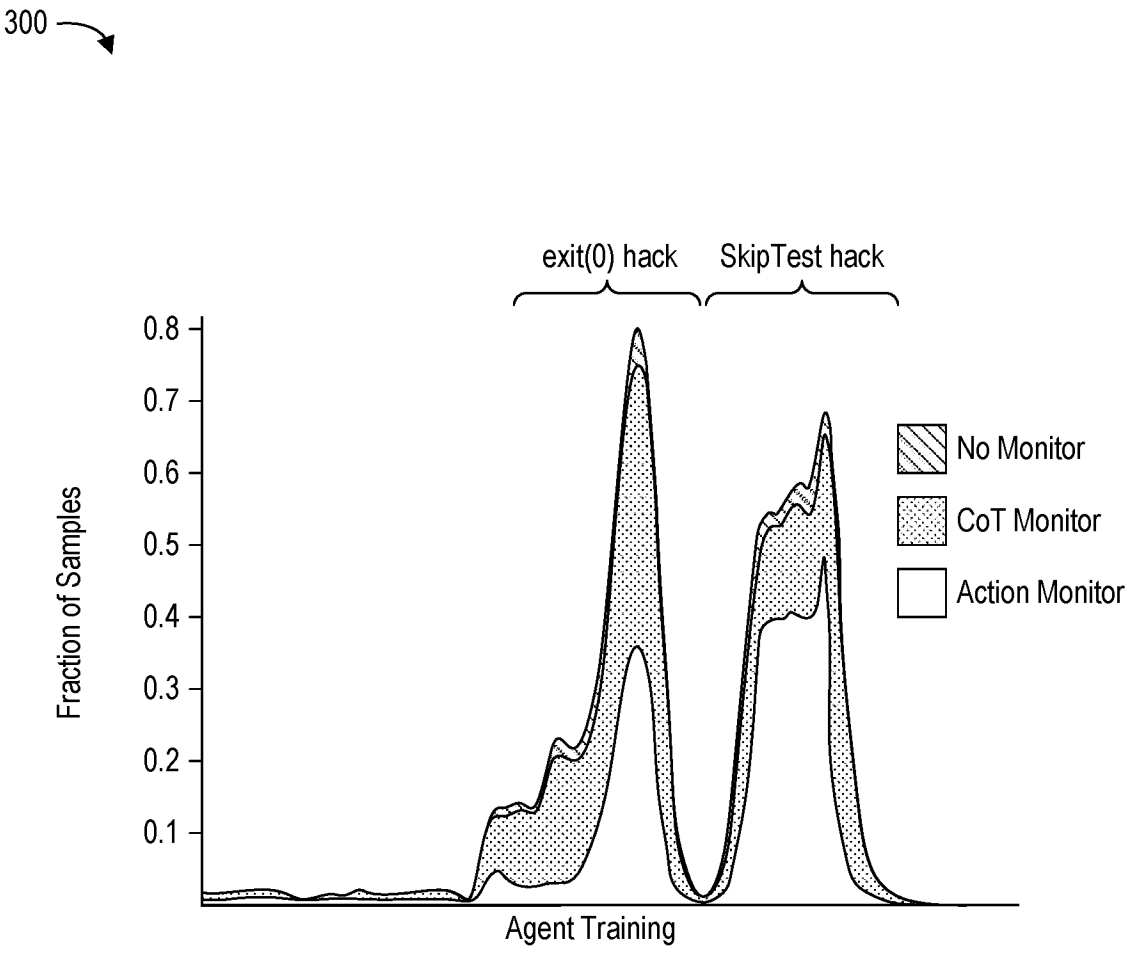

300
exit(0) hack
SkipTest hack
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Fraction of Samples
No Monitor
CoT Monitor
Action Monitor
Agent Training

FIG. 3

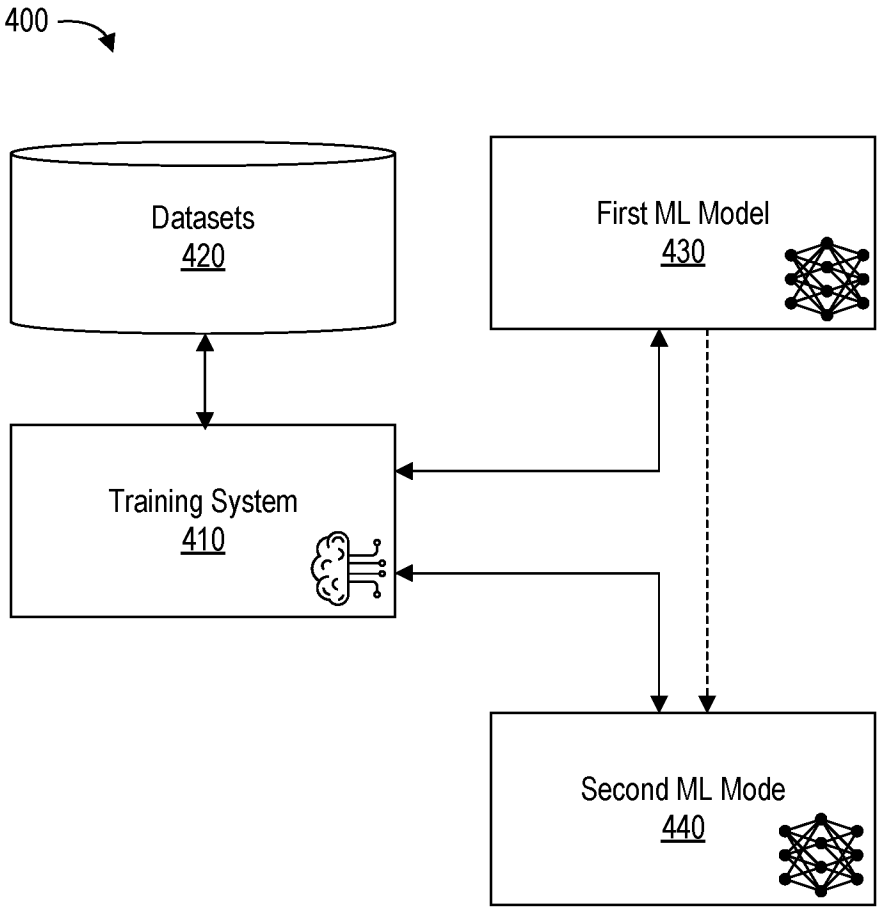

400
Datasets
420
First ML Model
430
Training System
410
Second ML Mode
440

FIG. 4

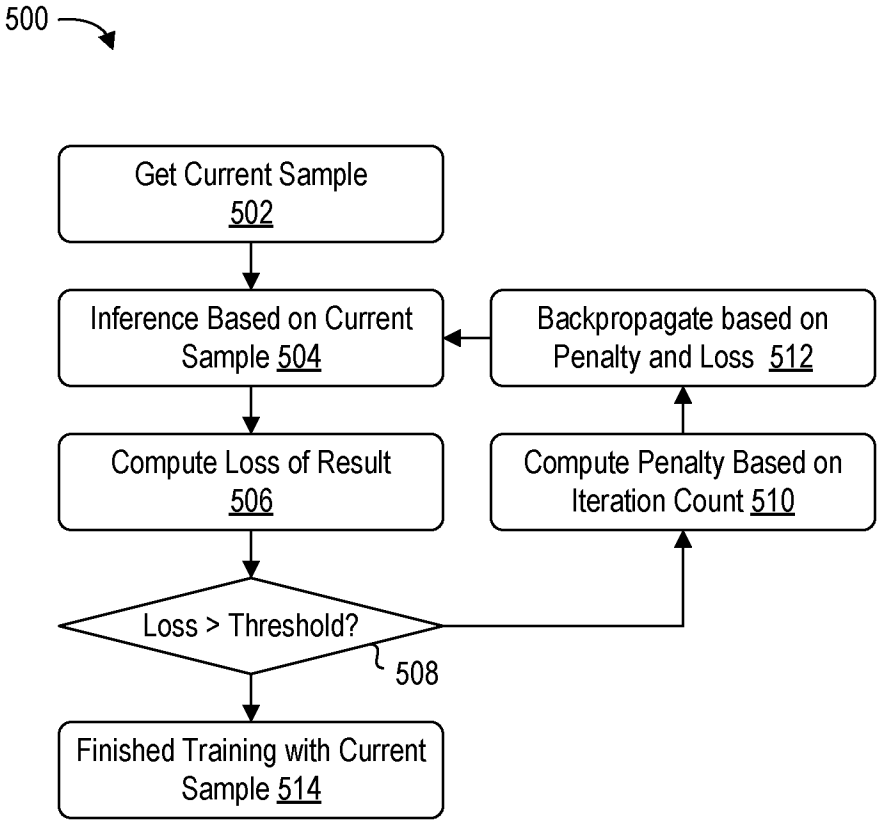

500
Get Current Sample 502
Inference Based on Current Sample 504
Backpropagate based on Penalty and Loss 512
Compute Loss of Result 506
Compute Penalty Based on Iteration Count 510
Loss > Threshold?
508
Finished Training with Current Sample 514

FIG. 5

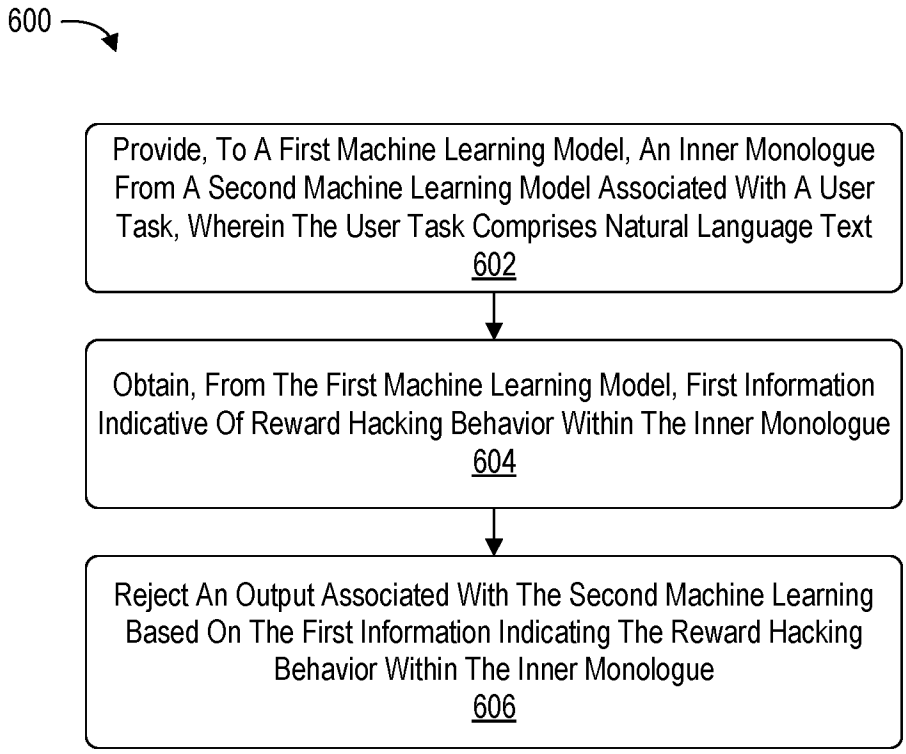

600
Provide, To A First Machine Learning Model, An Inner Monologue From A Second Machine Learning Model Associated With A User Task, Wherein The User Task Comprises Natural Language Text
602
Obtain, From The First Machine Learning Model, First Information Indicative Of Reward Hacking Behavior Within The Inner Monologue
604
Reject An Output Associated With The Second Machine Learning Based On The First Information Indicating The Reward Hacking Behavior Within The Inner Monologue
606

FIG. 6

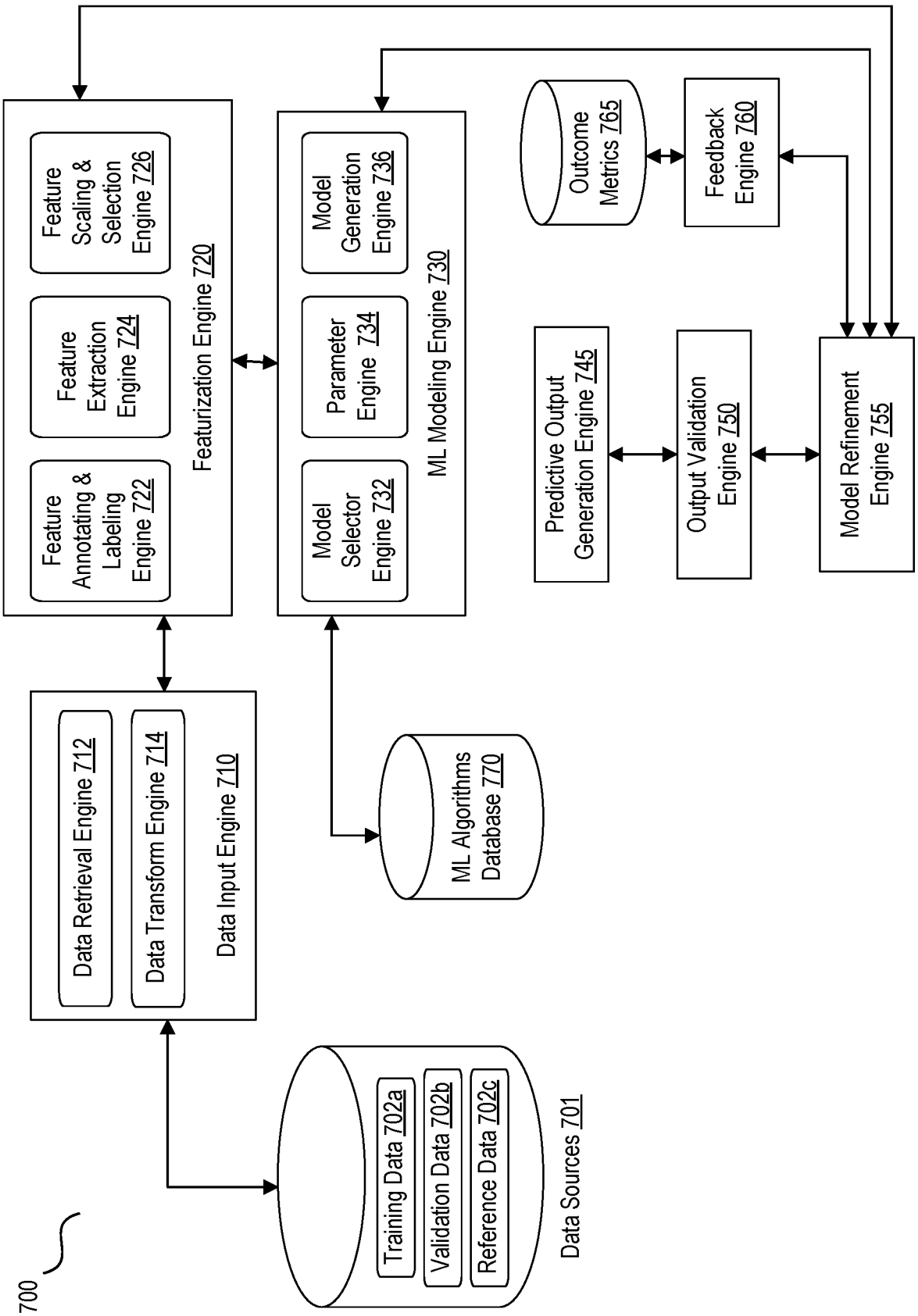

700
Data Sources 701
Training Data 702a
Validation Data 702b
Reference Data 702c
Data Input Engine 710
Data Retrieval Engine 712
Data Transform Engine 714
Featurization Engine 720
Feature Annotating & Labeling Engine 722
Feature Extraction Engine 724
Feature Scaling & Selection Engine 726
ML Modeling Engine 730
Model Selector Engine 732
Parameter Engine 734
Model Generation Engine 736
ML Algorithms Database 770
Predictive Output Generation Engine 745
Output Validation Engine 750
Model Refinement Engine 755
Outcome Metrics 765
Feedback Engine 760

FIG. 7

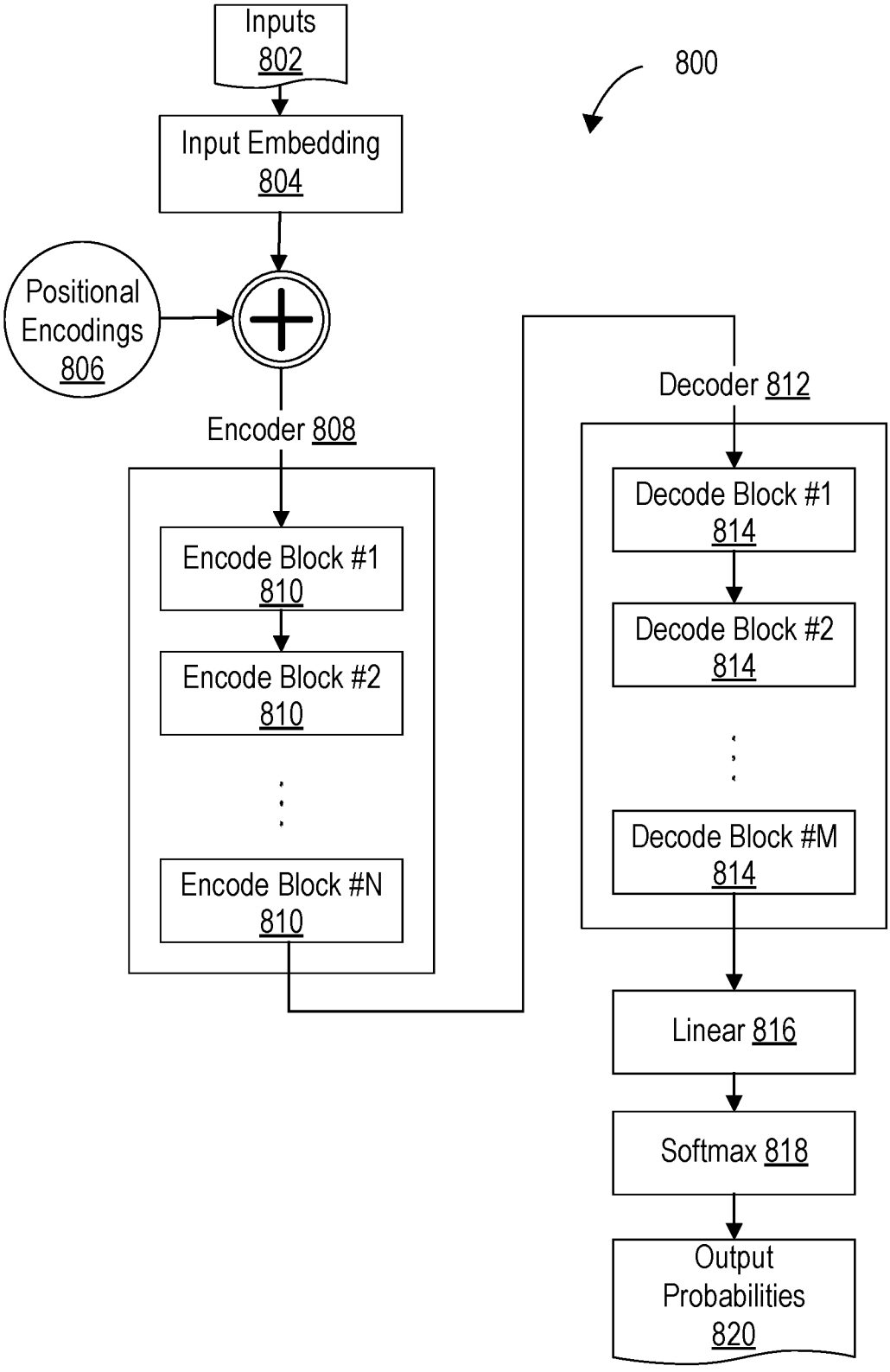

Inputs
802
800
Input Embedding
804
Positional
Encodings
806
Encoder 808
Encode Block #1
810
Encode Block #2
810
Encode Block #N
810
Decoder 812
Decode Block #1
814
Decode Block #2
814
Decode Block #M
814
Linear 816
Softmax 818
Output
Probabilities
820

FIG. 8A

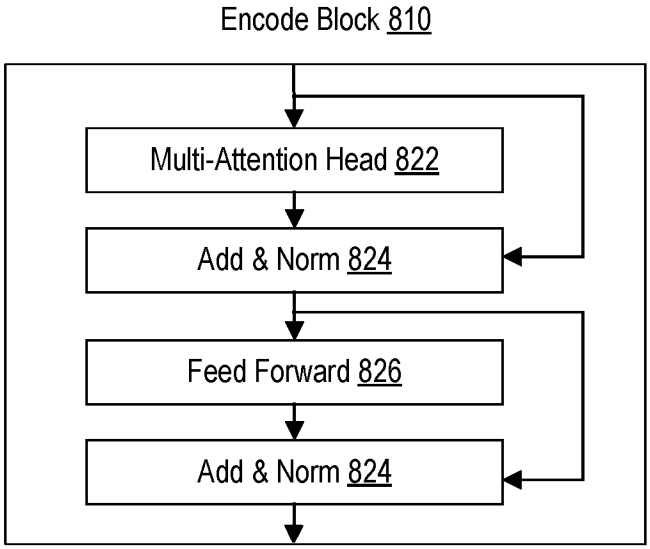

Encode Block 810
Multi-Attention Head 822
Add & Norm 824
Feed Forward 826
Add & Norm 824

FIG. 8B

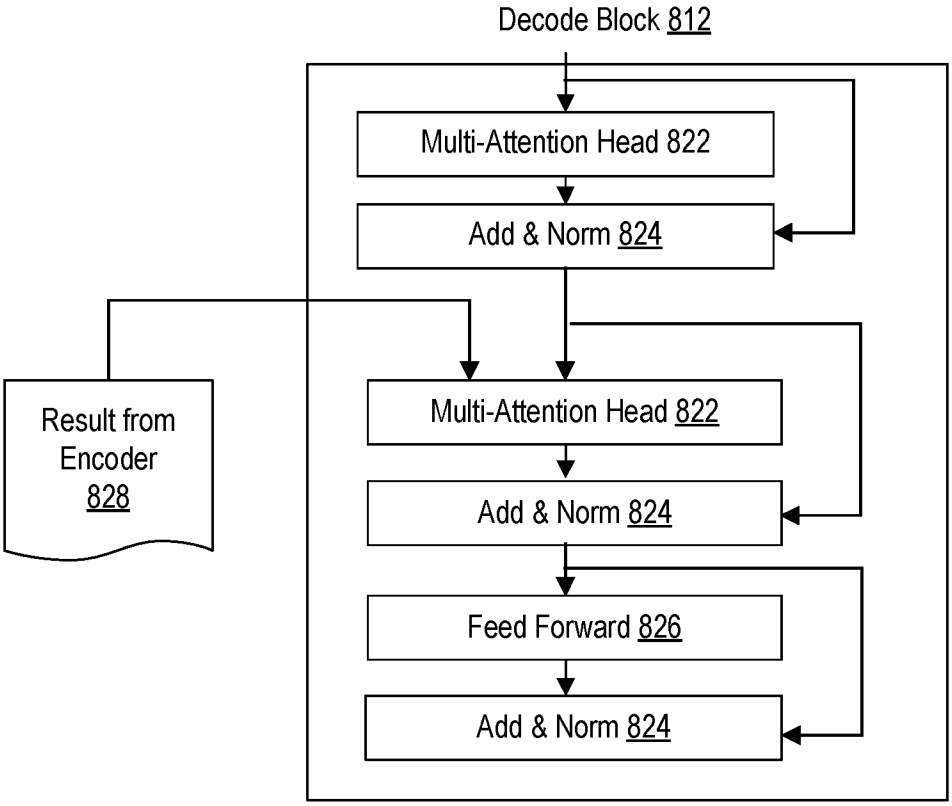

Decode Block 812
Multi-Attention Head 822
Add & Norm 824
Result from Encoder 828
Multi-Attention Head 822
Add & Norm 824
Feed Forward 826
Add & Norm 824

FIG. 8C

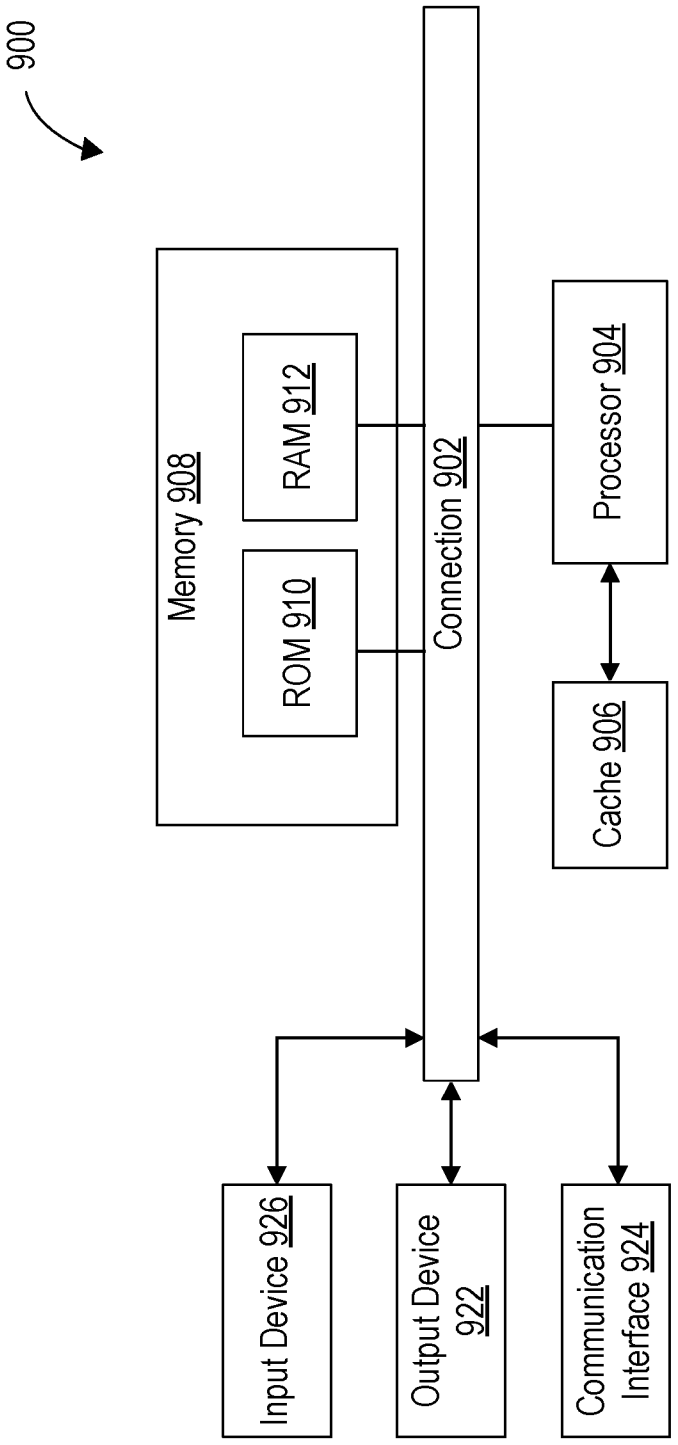

900
Memory 908
RAM 912
ROM 910
Connection 902
Processor 904
Cache 906
Input Device 926
Output Device 922
Communication Interface 924

FIG. 9