behavioral sciences

MDPI

Article

Automated Identification of Teacher Autonomy-Supportive Behaviors Based on Self-Determination Theory: A Deep Learning-Based Text Modeling Approach

Yatong Zu [1], Yu Feng [2], Tongyu Zhao [3], Hao Xu [3] and Jian Li [4,*]

[1] College of Physical Education, Jilin University, Changchun 130012, China; zuyatong@jlu.edu.cn
[2] Normal College, Yanbian University, Yanji 133002, China; yufeng3904@gmail.com
[3] College of Computer Science and Technology, Jilin University, Changchun 130012, China; id9269@alunos.uminho.pt (T.Z.); xuhao@jlu.edu.cn (H.X.)
[4] Tencent Technology Company Limited, Beijing 100193, China
* Correspondence: lijianjlu@126.com

Abstract

Self-determination theory (SDT) links classroom language with motivationally relevant instructional behaviors, but observing these behaviors at scale is resource intensive. Integrating natural language processing with traditional expert-based identification, this study extends established methods for identifying teachers' autonomy-supportive behavior and proposes an innovative, feasible SDT-based blended approach. This exploratory feasibility study tests whether SDT-informed verbal indicators of autonomy support and control can be computationally identified in authentic Chinese university teacher discourse. Eight lectures across eight disciplines provided 23.4 h of recordings and 5203 speech-derived teacher utterances. Two psychologists independently annotated them (Cohen's $\kappa = 0.818$); the 4904 agreed utterances form a 14-category consensus corpus that is 85.2% neutral teacher talk. In five-fold evaluation, a BERT–BiLSTM–CNN trained with an adaptive class-dependent cost (FOCC) reached a macro F1 of 30.2 ± 6.4% (95% CI 22.3–38.1), well above the imbalance-determined references of 6.6% and 7.1%. Differences among neural architectures and loss functions fell within fold variation. In a transparent character n-gram reference analysis, performance declined when contiguous corpus segments were held out, mainly because several rare behaviors clustered in particular segments. Corpus-level cue analysis showed that annotators coded instructional function rather than surface keywords. The findings establish the feasibility and current boundaries of SDT-informed discourse classification and position it as a tool for expert-assisted retrieval of classroom examples. Group-disjoint evaluation and validation against students' experiences are the next steps toward measurement claims.

Keywords: self-determination theory; teacher discourse; autonomy support; Chinese university classrooms; exploratory feasibility study; class imbalance; natural language processing

Academic Editors: Ines Testoni, Xiaochun Xie and Qiong Wu

Received: 27 September 2026
Revised: 4 November 2026
Accepted: 23 November 2026
Published: 27 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Learning motivation is an important foundation of learning behavior, and the relationship between motivation and behavior has long been a central concern in education (Lazowski & Hulleman, 2016). Self-determination theory (SDT) is a theory of

human motivation and behavior, and it is among the most effective motivational frameworks for interventions targeting learning motivation (Deci & Ryan, 1985). Applied to education, SDT addresses how to foster students' interest in learning, their valuing of education, and their confidence in their own capacities (Deci et al., 1991). Within its six mini-theories (Ryan & Deci, 2017, 2020), SDT describes a general sequence: the social context supports or frustrates the basic psychological needs for autonomy, competence, and relatedness; need satisfaction shapes the quality of motivation; and the quality of motivation shapes engagement, learning, and well-being (Niemiec & Ryan, 2009; Vansteenkiste et al., 2020). In the classroom, satisfying these needs creates a classroom climate that supports active learning, strengthens students' intrinsic motivation, and thereby shapes their learning behavior and academic engagement. SDT further emphasizes instructional conditions that support the internalization of extrinsic motivation, so that students' learning shifts from controlled toward autonomous regulation. External and introjected regulation are controlled forms of motivation, whereas identified and integrated regulation and intrinsic motivation are autonomous forms (Ryan & Deci, 2020). This distinction matters for learning: in a meta-analysis of 344 samples, intrinsic motivation and identified regulation were the most important predictors of students' adjustment, whereas external regulation was associated with lower well-being (Howard et al., 2021).

The theory also distinguishes need-supportive features of the social context, namely autonomy support, structure, and warmth, from need-thwarting features, namely control, chaos, and neglect (Vansteenkiste et al., 2020). Of these, autonomy support addresses most directly students' experience of their learning as self-endorsed: autonomy-supportive teachers attempt to understand and acknowledge students' perspectives and offer meaningful rationales and opportunities for input rather than relying on pressure (Reeve, 2009; Ryan & Deci, 2020). Such support is seen as promoting both autonomy and relatedness satisfaction and as nurturing intrinsic motivation and internalization, whereas external controls in the learning climate can undermine them (Deci et al., 1991; Niemiec & Ryan, 2009; Ryan & Deci, 2020). Within this framework, the teacher's interpersonal style is a central condition: whether instructional behavior is autonomy-supportive or controlling has a lasting influence on students' motivation and achievement (Reeve & Cheon, 2021). This view is consistent with the broader tradition of student-centered pedagogy, which asks teachers to create conditions for self-directed learning (Rogers & Freiberg, 1994), while SDT specifies the motivational processes through which such conditions take effect.

Empirical evidence supports these propositions across educational settings. A review of 51 autonomy-supportive teaching interventions, most of them randomized controlled trials, showed that teachers can learn to become more autonomy supportive and that this change benefits students, teachers, and the classroom climate (Reeve & Cheon, 2021). More broadly, a substantial body of research has linked autonomy-supportive teaching with engagement, persistence, perceived competence, creativity, and academic functioning (Boggiano et al., 1993; Grolnick & Ryan, 1987; Hardre & Reeve, 2003; Miserandino, 1996; Vallerand et al., 1997). In higher education, students who perceived their chemistry instructor as autonomy supportive showed increases in autonomous self-regulation, perceived competence, and enjoyment, together with reduced anxiety, over a semester (Black & Deci, 2000). The relevance of autonomy support is not limited to Western classrooms: among rural Chinese students, instructors' autonomy support predicted subsequent changes in autonomous motivation and perceived competence (M. Zhou et al., 2009). SDT also distinguishes autonomy support from structure. Clear expectations and guidance are positively related to autonomy support and contribute to students' engagement, so that structure and autonomy support are complementary rather

than opposing dimensions of teaching (Jang et al., 2010; Patzak & Zhang, 2025). This distinction is important for observation, because a directive or a time reference is not controlling by virtue of its form; its motivational meaning depends on its instructional function.

Among the social agents in students' learning, teachers hold a particularly proximal position: in a meta-analysis of the antecedents of student motivation, teacher autonomy support predicted need satisfaction and self-determined motivation more strongly than parental autonomy support did (Bureau et al., 2022). The outcomes associated with autonomy support are also broad. In a meta-analysis of need-supportive and need-thwarting behaviors in student populations (637 samples), autonomy support was positively correlated with engagement, the valuing of education, academic performance, and well-being, and negatively correlated with ill-being (Howard et al., 2025). These associations extend to higher education. Medical students who perceived their instructors as more autonomy supportive became more autonomous in their learning, which in turn accounted for increases in their perceived competence and in their endorsement of the psychosocial values taught in the course (Williams & Deci, 1996), and among Norwegian university biology students, perceptions of need-supportive lecturers were indirectly associated with higher achievement and lower dropout intentions through autonomous motivation and perceived competence (Jeno et al., 2018). Teacher autonomy support is therefore relevant not only to students' immediate classroom experience but also to the quality of their learning and their intentions to persist in their studies.

Teacher autonomy support is enacted through what teachers say and do during instruction (Reeve & Jang, 2006), and its effects can be traced to specific acts. In a study of 41 high school science classes, students' perceptions of particular daily supports, such as choice provision, rationales, and opportunities to ask questions, predicted increases in their engagement from one class session to the next (Patall et al., 2018). Likewise, autonomy-support training programs have been more effective when they trained multiple elements of autonomy support through skill-based activities (Su & Reeve, 2011). Improving autonomy-supportive teaching thus appears to depend on making its constituent behaviors concrete and recognizable. Identifying specific autonomy-supportive and controlling behaviors in authentic instruction is therefore a prerequisite for describing how autonomy support is enacted in particular classrooms, for giving teachers behavior-specific feedback, and for connecting SDT with everyday teaching practice.

Teachers organize instruction through dialogue, response patterns, and guidance, and their autonomy support or control is expressed in observable instructional behaviors. Drawing on SDT, educational researchers have mapped teachers' instructional behaviors onto three types: autonomy-supportive, controlling, and neutral (Reeve & Jang, 2006; Reeve et al., 2004). Established methods for assessing these behaviors include teacher self-reports, student ratings based on recall of teachers' behavior, and expert coding of classroom video recordings (Deci et al., 1981; Haerens et al., 2013; Reeve & Jang, 2006; Van den Berghe et al., 2013). These methods preserve classroom context and human interpretive judgment and have provided effective means of identifying and classifying teacher behavior.

Each of these methods, however, has limitations. Expert coding is labor intensive and difficult to repeat across many lessons. Reliable teacher-level scores from classroom observation typically require multiple lessons and multiple raters: for a mathematics observation instrument, three lessons each scored by two raters were identified as an optimal combination for research purposes, and developing reliable observation systems takes time, expertise, and considerable financial resources (Hill et al., 2012). Because students' experience of autonomy support varies across class sessions, with consequences

for their daily engagement (Patall et al., 2018), the few lessons that such designs can afford may capture a teacher's practice only partially. Teacher self-reports and student ratings are less costly but are subject to social desirability and recall bias, and the two perspectives do not necessarily converge: when mathematics teachers and their students rated the same classes, agreement was considerable for classroom management but low for cognitive autonomy (Kunter & Baumert, 2006). None of these methods readily captures the semantic dynamics of authentic classroom dialogue as it unfolds over time. These constraints limit the scale at which autonomy-supportive and controlling behaviors can be observed. A computational approach that complements expert judgment, rather than replacing it, is therefore needed to extend theory-informed observation to larger bodies of classroom discourse and to support teachers' reflection on their practice.

Research on automated classroom discourse analysis indicates that such an approach is feasible. Combining automatic speech recognition, natural language processing, and machine learning, Kelly et al. (2018) estimated the authenticity of teachers' questions in real classrooms, with computer-generated estimates correlating with human coding at the classroom level. Demszky et al. (2021) developed a computational measure of teachers' uptake of student contributions that was associated with instructional quality across several educational datasets. Utterance-level predictions have also been aggregated to estimate CLASS instructional support (Whitehill & LoCasale-Crouch, 2024), language models have been used to evaluate teachers' encouragement of student-to-student interaction (Ilagan et al., 2024) and classroom-management practices (Tan & Demszky, 2026), and multimodal features have been combined with ChatGPT-4 to assess encouragement and warmth (Hou et al., 2024). Automated measures can also inform professional learning: personalized, automated feedback on classroom discussion has been used to support teachers' reflection and instructional change (Jacobs et al., 2022). These studies, however, target question authenticity, uptake, global instructional quality, interactional facilitation, or classroom management. None of them operationalizes an SDT-informed taxonomy that distinguishes autonomy-supportive, controlling, and neutral teacher talk, and none uses naturally occurring Chinese university classroom discourse. The verbal expression of autonomy support therefore remains largely outside the scope of existing automated observation.

Natural language processing offers a means of addressing this gap. Classroom speech can be transcribed into text and analyzed at the level of individual teacher utterances, linking the instructional behaviors defined by SDT to their verbal expression in real classrooms. Teacher utterances can thus be assigned to finer categories, such as inviting students to speak, offering informational praise, or responding to student questions within autonomy support, so that autonomy-supportive, controlling, and neutral instructional behaviors can be distinguished. Such a connection cannot be reduced to generic text classification: the categories must be theoretically grounded, must preserve the distinction between support, control, and structure, and must remain interpretable to educational researchers and practitioners.

Against this background, the present study extends the identification of teacher autonomy-supportive behavior by integrating artificial intelligence and natural language processing with the traditional, expert-based approach. The resulting blended approach combines an SDT-informed observational taxonomy and independent expert annotation with computational classification of teacher discourse, so that theoretical definitions and expert judgment set the target and computational models extend its reach to authentic classroom speech. The study and its experiments are designed as an exploratory demonstration of the feasibility of this approach, examining how far it can identify the adapted categories in a strongly imbalanced corpus and where its present limits lie. The scope of identification is deliberately narrow. The approach targets verbal indicators of

autonomy-supportive and controlling behavior in teacher utterances from Chinese university lectures; it does not evaluate teaching effectiveness, and it does not measure students' motivation or need satisfaction. The study addresses three research questions:

RQ1: How can the autonomy-supportive instructional behaviors defined by SDT be translated into an observable classification scheme for teacher classroom discourse?

RQ2: To what extent can deep learning text models distinguish autonomy-supportive, controlling, and neutral instructional behaviors in authentic classroom transcripts?

RQ3: Which linguistic cues characterize autonomy-supportive and controlling teacher utterances in the annotated corpus, and how closely do these cues correspond to the coded instructional functions?

The study makes four contributions.

First, it contributes a novel SDT-based blended approach to identifying teacher autonomy-supportive behavior, which serves as the overarching contribution of the study. The approach integrates theory-guided behavioral operationalization, expert annotation, and computational classification into a single route from motivational theory to authentic classroom discourse. Rather than treating autonomy support only as an overall subjective perception or a post hoc rating, it identifies the specific verbal acts through which teachers enact autonomy support or control during classroom interaction. Its intended role is to assist expert evaluation, by retrieving concrete classroom examples for interpretation, and to support teachers' reflection on their practice.

Second, it provides empirical grounding for the observational operationalization of SDT. Trained psychology researchers independently annotated the teacher corpus using an SDT-based behavioral taxonomy, showing that autonomy-supportive teaching in authentic classrooms can be represented as a set of identifiable linguistic indicators.

Third, it builds a BERT–BiLSTM–CNN architecture that integrates contextual language representations, sequential features, and local phrase features to distinguish fine-grained instructional behaviors.

Fourth, it introduces the Fast Automatic Optimization of the Class-dependent Cost (FOCC) loss, which adaptively adjusts misclassification costs according to behavior categories and inter-class confusion. In this study, FOCC is an auxiliary methodological component: while the model learns from the full set of classroom utterances, FOCC helps it attend to the autonomy-supportive behaviors that are infrequent in the data but central to the measurement target.

2. Related Work

This section develops the rationale for the study in the following order. It first clarifies autonomy support, control, and structure as functional constructs, then reviews how teaching practice and autonomy support have been observed, with a focus on university classrooms, and what evidence higher education provides on their motivational relevance, and next sets out how observational categories can be translated into transcript indicators. It then examines automated analysis of teacher discourse, including large language models, and closes by identifying the gap that the blended approach addresses.

2.1. Autonomy Support, Control, and Structure as Functional Constructs

SDT distinguishes autonomy support and control by their functional significance for learners' volition rather than by their surface form. Informational feedback, perspective taking, and meaningful opportunities for input may support volition, whereas pressure and coercion may undermine it (Flink et al., 1990; Reeve et al., 1999). Reeve (2009) characterizes a controlling motivating style as the interpersonal sentiment and behavior through which teachers pressure students to think, feel, or behave in particular ways, and

he shows that teachers adopt this style partly in response to pressures from their own context, such as accountability demands; experimental evidence indicates that teachers who are held responsible for students meeting performance standards teach in a more controlling manner (Deci et al., 1982). The same analysis implies that controlling and autonomy-supportive behaviors are not fixed attributes of a teacher but recurrent choices in instruction, which makes them suitable targets for observation at the level of specific acts.

Recent work has refined this distinction in ways that are central to observation. First, autonomy support and structure are distinct and complementary dimensions, and their combination is associated with more favorable motivational outcomes (Patzak & Zhang, 2025), so a directive that provides needed guidance should not be equated with control. Second, the circumplex model of Aelterman et al. (2019) differentiates autonomy support, structure, control, and chaos into eight finer approaches, showing that these approaches relate to student and teacher outcomes in an ordered pattern; attuning and guiding approaches were associated with the most adaptive outcomes, and domineering and abandoning approaches with the least adaptive. Learning to adopt the student's perspective is a key route by which teachers move toward autonomy support (Reeve & Cheon, 2024). At the level of specific acts, an international Delphi study reached expert consensus on 57 teacher motivational behaviors consistent with SDT, providing shared labels and definitions for the instructional acts that interventions and observation schemes target (Ahmadi et al., 2023). Together, these studies establish that an observation scheme should code what a teacher's act does for the learner, not merely which words it contains, that it should separate supportive structure from control, and that it can be organized around discrete, nameable behaviors.

2.2. Observing Teaching and Autonomy Support in University Classrooms

Research on university teaching has shown that classroom practice can be observed systematically and that teacher talk is a central part of it. The Classroom Observation Protocol for Undergraduate STEM (COPUS) records instructor and student behaviors in 2 min intervals to characterize university classroom practices (Smith et al., 2013), and its application to 2008 classes taught by 548 faculty members at 25 North American institutions found that about 55% of the observed classes were didactic, with lecturing occupying 80% or more of class time (Stains et al., 2018). Teacher talk in university lectures can also be analyzed beyond its disciplinary content: in transcribed sessions of an introductory biology course at a large public university, noncontent instructor talk occurred in every analyzed session and could be reliably coded into categories such as building the instructor–student relationship and establishing classroom culture (Seidel et al., 2015). Teacher talk in university classrooms is therefore both pervasive and analyzable at the level of identifiable discourse moves.

Within SDT, Reeve and Jang (2006) videotaped one-on-one instructional sessions in which college students, randomly assigned to teacher and learner roles, worked on a learning activity; of 21 candidate teacher behaviors, eight were positively and six negatively related to learners' perceived autonomy. Observation of autonomy support in university classrooms has since been reported in a small number of studies. Using an observational checklist in university seminars, McLachlan and Hagger (2010) found that tutors who received brief training in autonomy-supportive teaching showed significant increases from baseline in two autonomy-supportive behaviors. In a 12-week intervention with university teachers in Colombia, perceived autonomy support increased and perceived controlling style decreased among students of trained teachers, whereas perceived autonomy support decreased in a comparison group, and expert coding of videotaped classes indicated that autonomy support predominated in more than 80% of

the trained teachers' interactions (Lozano-Jiménez et al., 2021). In a video-based study of 12 honors-program lessons taught by four teachers at a Dutch university of applied sciences, T. Kingma et al. (2024) coded 1509 teaching behaviors; structure-providing behaviors were the most frequent (46%), followed by teacher involvement (22%) and autonomy support (21%), whereas control accounted for 7%, with controlling language among its most frequent forms, and need-thwarting behaviors appeared only alongside need-supportive ones. Existing SDT-based observation instruments, most of them developed for school and physical education lessons (Haerens et al., 2013; Van den Berghe et al., 2013), combine durations, frequencies, and observable actions, and they preserve the contextual judgment that functional coding requires.

2.3. Autonomy Support in Higher Education

Evidence from higher education indicates that the motivational processes described above are not confined to school-age learners. A meta-analysis of higher-education studies found that perceived autonomy support was associated with students' autonomous motivation and need satisfaction with medium to large effects and with academic performance with small effects (Okada, 2023), and a systematic review of 17 studies in health professions education identified an autonomy-supportive learning climate as one of the interpersonal conditions that foster students' motivation (Orsini et al., 2016). Experimental evidence shows that the wording of instruction alone carries motivational meaning: when Belgian college students received a learning task introduced in autonomy-supportive language (e.g., "you can", "we suggest that you") rather than controlling language (e.g., "you should", "you have to"), they processed the material more deeply, performed better on a subsequent test, and persisted longer (Vansteenkiste et al., 2004). Longitudinal and cross-national studies point in the same direction. Among law students followed over three years, perceived faculty autonomy support predicted need satisfaction, which in turn predicted well-being and final grades (Sheldon & Krieger, 2007); in German and American university samples, informational feedback from professors was positively, and perceived pressure in the learning environment negatively, related to students' autonomy and competence (Levesque et al., 2004); in a Dutch higher-education sample, students who perceived high need support from their teachers showed relatively high autonomous motivation and achievement (Leenknecht et al., 2017); and among Norwegian university students in introductory calculus courses, perceived autonomy support predicted higher autonomous and lower controlled motivation, whereas perceived autonomy thwarting predicted the reverse pattern (Johansen et al., 2023).

Evidence from Chinese universities, the setting of the present study, is consistent with this pattern. Among 705 Chinese university students, perceived teacher autonomy support was positively related to need satisfaction, which together with autonomous motivation mediated its relation with class engagement (L. Zhou et al., 2022). Among Chinese university learners of English as a foreign language, perceived teacher autonomy support positively predicted agentic engagement through mastery-approach goals, whereas perceived controlling teaching negatively predicted it through performance-approach goals (Jiang & Zhang, 2021), and among engineering undergraduates in academic English speaking courses, perceived teacher autonomy support predicted speaking performance both directly and indirectly through basic psychological needs and classroom engagement (Y. Wang et al., 2024). Cross-cultural evidence further indicates that the benefits of need satisfaction are not specific to Western students: in university samples from Belgium, China, Peru, and the United States, the associations of need satisfaction with well-being and of need frustration with depressive symptoms were equivalent across countries (Chen et al., 2015).

On the teacher side, the circumplex of motivating and demotivating approaches has been replicated with higher-education teachers (Vermote et al., 2020), and among university teachers a self-reported teacher-centered approach was associated with more control and chaos, whereas a student-centered approach was associated with more autonomy support and structure (Huić et al., 2024). The lecture format itself does not settle the matter: undergraduates in lecture-based and problem-based programs did not differ in autonomous motivation, and controlling elements such as mandatory attendance were described as detrimental to motivation (Wijnia et al., 2011). These findings suggest that, in higher education as elsewhere, the motivational significance of instruction depends on how teachers communicate within a given format.

Two limitations of this literature motivate the present study. First, SDT-based observation instruments were developed mainly for laboratory tutoring sessions, school-age students, and physical education, where instruction is organized around practical tasks; in higher education, the few SDT-informed observational studies have involved small numbers of teachers (T. Kingma et al., 2024; Lozano-Jiménez et al., 2021; McLachlan & Hagger, 2010), and general-purpose protocols such as COPUS characterize instructional activities rather than autonomy support. University lectures, in which teacher exposition dominates and student participation is sparser, have therefore received far less SDT-informed observational attention, even though autonomy support is consequential in higher education (Black & Deci, 2000). The higher-education evidence reviewed above, moreover, rests largely on students' questionnaire reports, with the exception of experimental manipulations of instructional wording. Second, expert observation requires training and sustained coding effort, so it is rarely applied to large numbers of lessons. Questionnaires efficiently characterize students' perceptions but cannot locate the wording and timing of specific instructional acts. A method that identifies such acts in teacher language would complement both forms of evidence, provided that its categories remain faithful to the functional definitions above.

2.4. From Observational Categories to Transcript Indicators

Translating an observational scheme into transcript categories requires decisions about which behaviors have a verbal realization. The candidate scheme of Reeve and Jang (2006) already included deadline statements, which were too infrequent for their subsequent analyses, and it combined durations, frequencies, and observable actions. The present single-utterance labels are therefore adaptations of selected categories rather than a modality-equivalent reproduction, and Table 1 makes the differences explicit.

Table 1. Adaptations from observation to transcript coding.

Source Behavior or Feature	Treatment in This Study	Construct Boundary
Listening; holding or monopolizing materials	Excluded	Silence and physical control are assessed through audiovisual evidence
Allowing students to work in their own way	Retained through its verbal realization, coded as ask-stu	The duration of independent work requires observation of the learning activity; the transcript captures the teacher's statement or question that leaves the choice to students
Student talking	Retained label stu-tk for teacher invitations to talk	Identifies an invitation to participate; actual student speech duration is a complementary indicator
Informational praise; encouragement; hints; perspective taking; responsiveness	Retained verbal categories	Their function may require dialogue, tone, and student response

Asking what students want to work on	Coded together with verbal offers of choice as ask-stu	Questions about students' preferences and statements granting choice share the same supportive function
Uttering answers; directives; should or ought statements; controlling questions	Retained as provisional verbal indicators	Task context helps distinguish supportive structure from pressure
Exhibiting solutions or answers	Retained as a provisional transcript category	Verbal and visual evidence together clarify the distinction from uttering answers
Deadline statements	Retained from the broader source scheme	Interpretation distinguishes organizational time notices from coercive pressure
Neutral teacher talk	Added residual text class	Residual category for segments without an identified support or control indicator

Four principles guided these adaptations. The first is channel correspondence: behaviors whose evidence is non-verbal, such as listening, silence, or physical control of learning materials, were excluded because a transcript cannot represent them without inference beyond the data. The second is unit correspondence: behaviors that the source scheme quantifies by duration, such as time allowing students to work in their own way and time students spend talking, do not map onto a single utterance. For these behaviors, the codebook captures the teacher's verbal realization, namely the statement or question that leaves a choice to students (ask-stu) or invites them to talk (stu-tk), whereas the duration itself would require interval coding of the learning activity. The third is verbal realization: behaviors with a characteristic verbal form were retained even when their source definition also includes non-verbal aspects. In the corpus, ex-ans utterances frequently accompany the display of a worked solution or answer and direct students' attention to it (for example, 我给大家看, "let me show you", or 大家看啊，你看到这个图, "everyone look, you see this figure"), whereas ut-ans utterances often pose a question and supply the answer at once (the trigram 是什么, "what is", occurs in 27% of ut-ans utterances). The fourth is exhaustive coverage: a residual neutral category was added because continuous lecture talk, unlike a set of pre-selected behavioral episodes, requires every segment to receive a label.

Deadline statements show how these principles bear on construct validity. Reeve and Jang (2006) listed deadline statements among their candidate controlling behaviors, and imposed time limits are a recognized source of pressure within SDT. A reference to time can, however, also provide structure. The codebook therefore assigned ddl only when a time reference functioned as a demand on students' activity, and coded organizational or narrative time references as neutral teacher talk. The agreed corpus shows that this boundary was applied in practice: of the 83 utterances containing 分钟 ("minute"), 38 were coded as deadline statements and 35 as neutral teacher talk, the latter mostly describing durations as lecture content (Section 4.5). The adapted scheme is therefore a transcript-specific operationalization derived from the SDT observational tradition rather than a replication of the original instrument. Its construct validity rests on an explicit correspondence between each retained category and a verbal function, which Tables 1 and 2 and the corpus evidence reported in Sections 4.5 and 4.6 make open to inspection.

Table 2. Behavioral labels and transcript interpretation boundaries.

Code	Behavioral Indicator	Definition and Boundary	Oriennation	*n*
tch-tk	Teacher talk	Content explanation without an identified support or control indicator, such as an explanation of water molecules. Neutral refers to the residual codebook category.	Neutral	4180

ut-ans	Uttering solutions or answers	Revealing a solution before students can discover it; interpretation considers the opportunity for independent discovery.	Control	67
ut-comd	Uttering directives or commands	Directives such as “Put it on its side.” Their instructional function is interpreted with task context.	Control	110
mk-sud	Making should or ought statements	Expressions using should, must, or ought; their function depends on the demand and surrounding interaction.	Control	29
stu-tk	Teacher invitation to student talk	Invitations to discuss or speak, representing a teacher-language indicator of participation opportunity.	Support	42
ddl	Deadline statements	Expressions concerning time limits or remaining time; context helps distinguish pressure from organization.	Control	64
praise	Praise as informational feedback	Feedback about improvement or mastery; the interpretation of positive approval considers its informational function.	Support	14
ask-stu	Offering choice or asking what students want	Statements or questions that leave what or how to work to students or ask about their interests, such as 你来决定 (“you decide”) or 大家可以随便说出来 (“feel free to say it”).	Support	17
comm	Perspective taking statements	Acknowledging students’ experience, such as acknowledging difficulty.	Support	35
hints	Offering hints	Suggestions preserving a route to discovery; the difference from giving an answer depends on task context.	Support	245
ex-ans	Exhibiting solutions or answers	The retained ex-ans category concerns exhibiting a solution or answer; visual and task context help distinguish it from ut-ans.	Control	17
con-q	Asking controlling questions	Directives phrased as questions; interpretation considers the degree of choice and pressure.	Control	40
offer-e	Offering encouragement	Expressions encouraging engagement, such as “You can do it”; their function is read with the learner’s situation.	Support	18
resp	Responding to student questions	Responses to student contributions; preceding turns help establish the contingent relationship.	Support	26

2.5. Automated Analysis of Teacher Discourse

The automated studies introduced in Section 1 can be compared along four dimensions that define the present design space. The first is the observation target. Existing systems measure question authenticity (Kelly et al., 2018), uptake of student contributions (Demszky et al., 2021), global instructional support (Whitehill & LoCasale-Crouch, 2024), encouragement of student-to-student interaction in a simulated classroom discussion (Ilagan et al., 2024), classroom management (Tan & Demszky, 2026), or encouragement and warmth (Hou et al., 2024). Each target is derived from its own framework, and none corresponds to the SDT distinction between autonomy-supportive, controlling, and neutral talk. A systematic review of 68 studies on artificial intelligence in classroom discourse also found that most were conducted in primary and secondary science and language classrooms (D. Wang et al., 2024), which suggests that university classrooms remain comparatively underrepresented. The second dimension is the unit of analysis: most systems classify utterances or turns and then aggregate predictions to lessons, as Whitehill and LoCasale-Crouch (2024) did for CLASS scores, so utterance-level identification is an established building block for lesson-level description.

The third dimension is the data condition. Systems that operate on recorded speech must cope with recognition errors, and Kelly et al. (2018) showed that classroom-level estimates from automatically recognized audio can still agree with human coding. Behaviors of theoretical interest are also typically rare relative to ordinary exposition, so the learning problem is strongly imbalanced; loss functions that down-weight easy

examples or assign class-dependent costs have been developed for such conditions (Khan et al., 2018; Lin et al., 2017), and contextual encoders such as BERT provide representations that are sensitive to the wording surrounding a salient expression (Devlin et al., 2018). The fourth dimension is the intended use. Automated feedback has been used to prompt teachers' reflection on their discourse (Jacobs et al., 2022). In a randomized controlled trial with 1136 instructors in a large online computer science course, feedback from a natural language processing tool increased instructors' uptake of student contributions by 13% (Demszky et al., 2024), whereas a study of the Teacher Talk Tool found that most feedback was perceived as accurate and prompted a high degree of reflection, while evidence that teachers took up the feedback was more limited (Kelly et al., 2025). These findings suggest that automated discourse measures can inform instructional practice, and they support a design in which automated output supports human interpretation rather than replacing it.

Viewed along these dimensions, the present study remedies limitations of prior research: an SDT-derived target, utterance-level identification of verbal acts, speech-derived Chinese text from university lectures, a strongly imbalanced category distribution, and an expert-assisted rather than evaluative use.

2.6. Large Language Models as an Alternative

Instruction-tuned large language models (LLMs) raise the question of whether such indicators could be identified without task-specific training, through zero-shot or few-shot prompting. This possibility is attractive for low-frequency categories, where supervised models have few examples. Evidence from adjacent tasks, however, indicates reliability concerns that are directly relevant to psychologically meaningful labels. Hang et al. (2025) found that LLMs used for scientific literature discovery produce hallucinated references and that citation recall varies with model and retrieval configuration, and Dai et al. (2025) showed that LLM-generated citations reflect source-related bias independent of content. Although these studies concern literature discovery and citation rather than classroom discourse, they show that generated evidence can be fluent yet unsupported, and that the selection of evidence can be systematically biased. For classroom classification, analogous risks include labels that vary across prompts or runs, explanations that cite wording not present in the utterance, and interpretations shaped by the model's prior associations of particular teaching styles with support or control.

Studies in classroom settings indicate both the potential and the limits of LLMs. In Chinese middle-school mathematics and Chinese-language lessons, a customized GPT-4 model coded classroom dialogue with high consistency with human coders and with considerable time savings (Long et al., 2024). In an evaluation of eight LLM endpoints against certified expert ratings of primary classroom lessons, however, Melo et al. (2026) found substantial variation across repeated runs, at best moderate agreement with the experts, and a tendency to justify judgments with explicit verbal cues rather than contextual evidence. Hou et al. (2024) used ChatGPT for zero-shot assessment of encouragement and warmth within a multimodal pipeline. The work closest to the present task concerns teachers' motivational messages. Metzner et al. (2025) fine-tuned an open-weight LLM to classify SDT-based need-supportive and need-thwarting messages in transcribed lessons taught by 137 pre-service teachers in German schools; supportive messages reached F1 values of 0.73–0.80, whereas the less frequent thwarting messages proved considerably harder (F1 0.27–0.61), even though synthetic examples were added to offset the imbalance. A subsequent review characterized LLM-based classification of teachers' motivational messages as a promising but still emerging field, with evidence drawn so far from German and Spanish schools and with open questions concerning imbalanced categories, sentence-level annotation without surrounding discourse, and

generalizability beyond these contexts (Metzner et al., 2026). These studies show that LLMs can contribute to classroom observation, and they also emphasize the importance of a clearly specified observation target and of human verification.

To our knowledge, however, no study has yet applied LLMs to identify observation-derived autonomy-supportive and controlling teacher behaviors in Chinese-language university lectures, a setting in which teacher exposition dominates and the behaviors of interest are rare. A credible LLM comparison would therefore apply the same codebook and the same held-out partitions, draw few-shot examples only from training data, constrain outputs to the label set, record prompts and model versions, repeat runs to estimate label consistency, and verify any cited evidence against the utterance. The supervised setting reported here provides the reference point for such a comparison: its category supports, imbalance-determined references, and confusion patterns define what an LLM would need to exceed to be useful for this task.

2.7. Synthesis and Research Gap

The literature reviewed above narrows to a specific gap. SDT defines autonomy support and control by their function for learners and separates them from structure (Section 2.1). Observational research, including a small number of studies in university classrooms, has shown that these functions can be coded systematically, but SDT-based instruments were developed mainly for school-age and physical education settings, and their cost limits scale (Section 2.2). Research in higher education, including Chinese universities, indicates that autonomy support and control are associated with university students' motivation, engagement, and achievement, but this evidence rests mainly on questionnaires (Section 2.3). Part of the observational repertoire has a verbal realization that can be represented in transcripts, provided that the adaptation is explicit (Section 2.4). Automated discourse analysis has demonstrated that teacher talk can be classified at the utterance level from speech-derived text, but for targets other than SDT-defined autonomy support and not in Chinese university classrooms (Section 2.5), and LLM-based approaches, although emerging for related categories of teacher messages, have not yet been applied to this target and setting and still face unresolved reliability concerns (Section 2.6). The present study addresses this gap with a blended approach in which the SDT-informed taxonomy and independent expert annotation define the target, and computational classification extends its identification to authentic teacher discourse, with its scope restricted to verbal indicators rather than motivational effects.

3. Research Design and Methods

3.1. Theoretical Rationale and Exploratory Design

The study uses an exploratory, observational, single-label design. RQ1 concerns the translation of theoretical categories into verbal indicators and the empirical relation between wording and coded function; RQ2 concerns classification against expert consensus labels relative to imbalance-determined references; and RQ3 concerns corpus-level and illustrative linguistic cues. The analysis unit is a speech-derived teacher segment at approximately sentence level. Generalization across classrooms and teachers is treated as a separate question from this utterance-level task, and the evidence relevant to it is reported in Section 4.3.

Each retained segment receives the label that most directly represents its dominant coded indicator. The original annotation description identifies two instances of simultaneous verbal behaviors, supporting the practical use of a single-label task in this material. At the interactional level, supportive and controlling practices may nevertheless coexist across a longer exchange.

The adapted scheme contains seven support-related, six control-related, and one neutral label. These orientations organize theoretically relevant instructional functions within the verbal codebook. Table 1 explains the adaptation from observation to transcript analysis, and Table 2 provides the final category definitions and supports.

3.2. Participants, Classroom Corpus, and Transcript Preparation

The corpus was developed from audio recordings of eight in-person university lectures in China, covering computer science, mathematics, psychology, medicine, nursing, environmental studies, finance, and philosophy. The lectures were delivered by seven university teachers, all male, with a mean age of 43.7 years. All recordings were made in lecture-based classes in which teacher exposition was the dominant mode of instruction, and together they contained 23.4 h of teaching. A microphone was used to capture only the teacher's speech; no student speech or other student information was recorded. The participating teachers were fully informed of the entire study procedure and signed informed consent forms before recording.

Teacher audio was divided into segments at natural speech pauses using automatic audio segmentation based on the generalized likelihood ratio (D. Wang et al., 2008), and each segment was converted into Chinese text by automatic speech recognition (ASR; see Malik et al., 2021, for an overview of the technology). The resulting material comprised 5203 teacher utterances at approximately sentence level. Transcript errors were retained because the goal was to examine an approach that operates on speech-derived text from authentic classrooms rather than on idealized manual transcripts. Across the 5203 utterances, the longest contained 107 Chinese characters, the shortest five, and the mean length was 55.32 characters; in the 4904-utterance consensus corpus the mean length is 55.03 characters.

The seven male teachers in the sample taught eight university subjects. This breadth introduces variation in disciplinary vocabulary and instructional language, providing a useful setting for initial corpus development. Interpretation remains anchored in these eight instructional settings.

Retaining ASR output places the task closer to the input conditions of classroom recording workflows, but it involves a trade-off: some categories depend on short function words or on very few words in total, and recognition errors affecting these elements could alter the instructional meaning of a segment. Table 3 quantifies this exposure in the consensus corpus. Every should or ought statement contains a modal marker, and 88.1% of ut-ans and 87.5% of con-q utterances contain a question particle or question mark, so the omission of a modal or question particle could turn a suggestion into an apparent directive, or an invitation into a statement. Praise, encouragement, and responsiveness are also the shortest categories, with means of 26.0–28.1 characters against 54.7 for neutral talk, so a single misrecognized word represents a larger share of the available evidence. These categories are therefore expected to be most sensitive to transcription quality, and a comparison with manually corrected transcripts for them is a priority for subsequent validation.

Table 3. Utterance length and selected linguistic markers by category in the consensus corpus.

Label	*n*	Mean Length (Characters)	Modal Marker (%)	Question Particle or Mark (%)	Time Expression (%)
tch-tk	4180	54.7	42.3	48.5	14.2
ut-ans	67	58.0	38.8	88.1	10.4
ut-comd	110	58.9	70.0	47.3	5.5
mk-sud	29	61.8	100.0	69.0	13.8
stu-tk	42	60.7	45.2	50.0	23.8
ddl	64	48.9	31.2	51.6	85.9

praise	14	27.2	71.4	7.1	0.0
ask-stu	17	64.9	47.1	47.1	11.8
comm	35	52.7	45.7	40.0	14.3
hints	245	62.0	50.2	53.9	13.1
ex-ans	17	60.3	29.4	52.9	23.5
con-q	40	62.6	40.0	87.5	17.5
offer-e	18	28.1	22.2	22.2	0.0
resp	26	26	23.8	11.5	23.1

Note. Markers are detected by string matching in the agreed ASR text. Modal markers: 应该, 必须, 要, 得, 需要, 一定; question markers: 吗, 呢, 吧, and question marks; time expressions: 分钟, 秒, 时间, 还有, 最后, 截止, 小时. Several markers are polysemous, so the percentages describe exposure to these forms rather than their instructional function.

3.3. Coding and Annotation

Two trained psychologists with expertise in autonomy support independently annotated the 5203 utterances using the 14-label scheme. Overall agreement was Cohen's $\kappa = 0.818$ (Cohen, 1960), with identical labels for 4904 cases, or 94.25% of the material. Given this observed agreement, the reported κ implies a chance agreement of approximately 0.684, which is consistent with the strong predominance of neutral teacher talk; κ is therefore the more informative of the two figures. The agreed cases form the modeling corpus, and the remaining 299 cases (5.75%) were set aside. Agreement on a specific behavioral label also implies agreement on its autonomy-supportive, controlling, or neutral orientation within the retained corpus, so the corpus provides a common expert reference at both levels.

Table 2 and Figure 1 describe the same 4904-utterance consensus corpus. Mapping the agreed labels to their orientations yields 397 support-related, 327 control-related, and 4180 neutral utterances. All figures and tables are derived from this consensus data file. Annotator-level marginals are not reported, because they also include the utterances on which the two annotators disagreed and therefore do not describe the modeling corpus. The hierarchy distinguishes orientation from the specific verbal act: two annotators may agree on a supportive orientation while distinguishing differently between a hint and encouragement, and directives and controlling questions illustrate an analogous distinction within control-related language. Neutral teacher talk is represented by a single residual label whose coding boundary concerns whether a specific support- or control-related indicator is present.

The exclusion of the 299 non-agreed cases has a bounded effect on the corpus composition. If all of them were re-included, the share of neutral teacher talk would lie between 80.3% (if none were neutral) and 86.1% (if all were neutral), compared with 85.2% in the consensus corpus. The consensus corpus therefore does not substantially misrepresent the imbalance profile of the full material. The excluded cases are nonetheless likely to be enriched for boundary cases, such as segments in which a hint and a supplied answer, or an invitation and a controlling question, are difficult to separate from the transcript alone. The consensus labels thus define a clear reference for the present feasibility evaluation, and adjudication of the excluded cases is identified in Section 5.6 as a priority for extending coverage of ambiguous discourse.

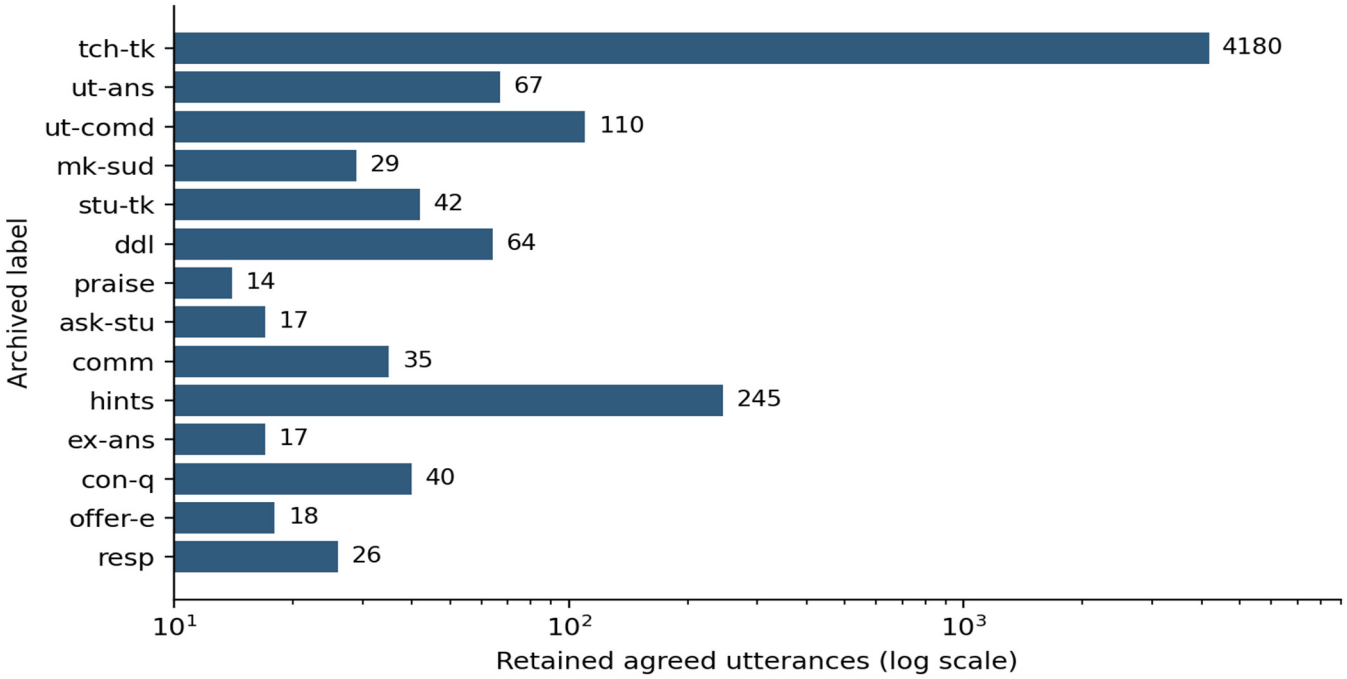


Figure 1. Category distribution in the 4904-utterance consensus corpus. Counts correspond to Table 2; the logarithmic axis makes the rare categories visible.

3.4. Deep Learning Text Model

The classifier receives one Chinese teacher segment. The encoder is a 12-layer Chinese BERT-Base model pretrained on Chinese Wikipedia (Devlin et al., 2018), with 768-dimensional token representations at each layer. Concatenating the final four layers gives $n \times 3072$ features, where n is the tokenized sequence length. A two-layer BiLSTM with hidden width 256 produces an $n \times 512$ bidirectional representation. Concatenating the BERT and BiLSTM outputs, as a residual connection, gives $n \times 3584$ features. Three parallel CNN branches use kernel widths 2, 3, and 4, with 256 filters per branch, and their pooled outputs form a 768-dimensional feature vector for a fully connected 14-label classification layer.

This architecture combines contextual token representations, bidirectional sequence information, and local phrase features. Its rationale is that instructional meaning can depend on both a salient expression and the wording surrounding it. For example, an evaluative word gains meaning from the behavior being evaluated, and a time expression gains instructional significance from the accompanying task demand. Figure 2 summarizes the architecture, and Table 4 lists the training settings.

Table 4. Architecture and training settings.

Component	Reported Settings
Input	Chinese ASR-derived teacher segments; 5–107 characters (mean 55.32 over 5203 utterances; 55.03 over the 4904 agreed utterances); segmentation by generalized likelihood ratio at speech pauses
Encoder and classifier	Chinese BERT-Base, 12 layers, width 768, pretrained on Chinese Wikipedia; final four layers concatenated to 3072. Two-layer BiLSTM: hidden width 256, bidirectional output 512. Residual concatenation: 3584. Three CNN branches: widths 2, 3, 4; 256 filters each. Final feature width: 768; fully connected 14-label output.
Neural training	Adam; fixed 20 epochs; batch size 8; learning rate 5×10^{-5}; seed 42; PyTorch 1.13.1; RTX 3090 GPU; i5-12400 CPU

Evaluation	Five folds over utterances; 5% of each training portion held out for validation; accuracy and macro F1 as mean ± fold SD with t-based 95% CIs; imbalance-determined references (always-neutral, prior-matched guessing)
FOCC configuration	$\gamma = 3$ ($\gamma = 1$–6 examined); diagonal costs initialized to 1 and off-diagonal costs to 30; cost learning rate 0.1; classifier updates by batch and cost updates by epoch; class cost read from the off-diagonal row mean (Equation (6))

Context in this model refers to wording within the teacher segment. Surrounding dialogue, prosody, and visual behavior provide complementary evidence for interpreting an utterance's interactional function.

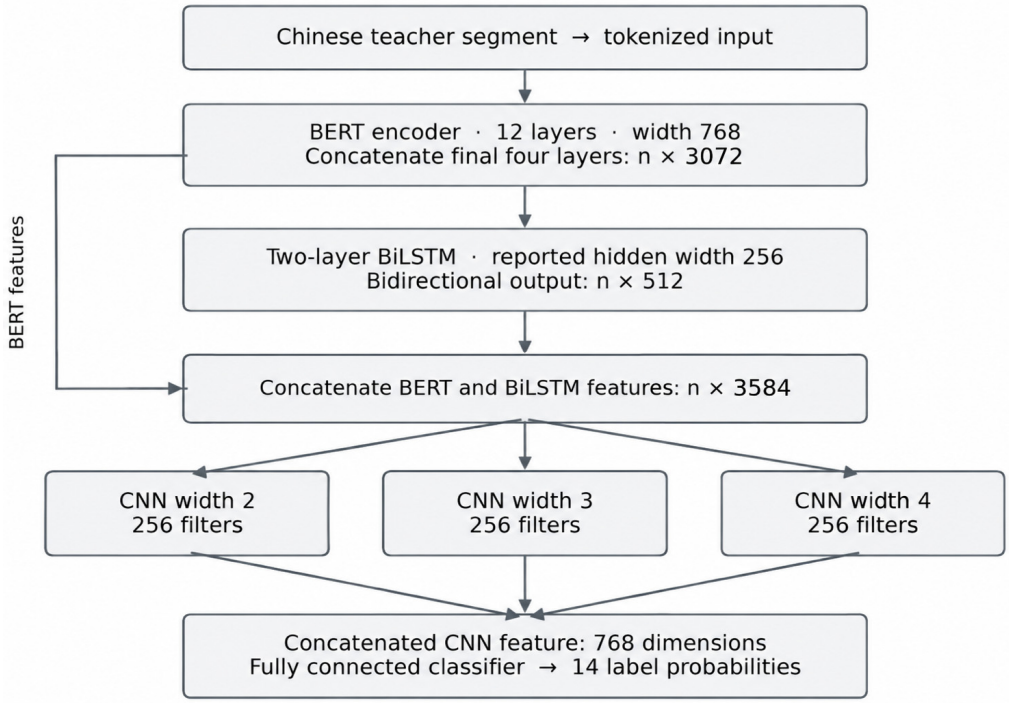


Figure 2. BERT–BiLSTM–CNN architecture with the dimensions reported for each feature transformation. The diagram summarizes the classifier pathway; the adaptive-cost configuration is described separately.

3.5. Class Imbalance and the Training Objective

The strong predominance of neutral teacher talk motivates evaluating losses that alter the contribution of examples to training. Focal loss reduces the contribution of examples already assigned high probability (Lin et al., 2017), while cost-sensitive learning introduces differing error penalties. Khan et al. (2018) provide the background for learned class-dependent costs, and Ashrafi et al. (2020) illustrate cost-sensitive contextual modeling in another language task. These approaches offer complementary ways to investigate learning under imbalance.

The Fast Automatic Optimization of the Class-dependent Cost (FOCC) configuration combines focal modulation with class-dependent costs that are learned during training. Two objectives are involved: one updates the network parameters θ after each batch, and one updates a C × C cost matrix ξ after each epoch. The classifier-level objective is a cost-weighted focal loss (Equation (1)).

$$L_\theta = -\frac{1}{N}\sum_{r=1}^{N} w_{y_r}\,(1 - p_{r,y_r})^{\gamma} \log p_{r,y_r} \quad (1)$$

Here y_r is the integer reference label of example r, p_{r,y_r} is the corresponding predicted probability, w_{y_r} is the scalar cost of the reference class, N is the batch size, and γ is the focal

parameter. The cost matrix is updated from three sources of information identified in earlier cost-sensitive work (Khan et al., 2018): class distribution, class separability, and current classification errors. The class-distribution matrix D is defined in Equation (2), where n_i is the number of training examples of class i.

$$D_{ij} = \max(n_i, n_j) \tag{2}$$

Class separability is computed from the difference between the probability assigned to an example's reference class i and the probability assigned to another class j, averaged over the N_i training examples of class i (Equation (3)). A small difference indicates poor separability and yields a large S_{ij}. Written in its numerically stabilized form, the expression includes a small constant $\varepsilon > 0$ inside the logarithm, so that it remains finite when the two probabilities are nearly identical.

$$S_{ij} = -\frac{1}{N_i}\sum_{k=1}^{N_i} \log\left[\left(p_{k,i} - p_{k,j}\right)^2 + \varepsilon\right] \tag{3}$$

The target cost matrix M combines these components with R, the confusion matrix on the validation portion, using element-wise products and element-wise exponentials (Equation (4)). The cost matrix is then updated by minimizing the objective in Equation (5).

$$M = D \odot \exp(-S) \odot \exp(-R) \tag{4}$$

$$L_\xi = \|M - \xi\|_F^2 + \mathrm{CE}_{\mathrm{val}}(\theta) \tag{5}$$

Because θ is held fixed during the cost update, the validation cross-entropy term in Equation (5) is constant with respect to ξ, and the gradient with respect to the cost matrix is $-2(M - \xi)$. The validation evidence therefore affects ξ through R inside M, whereas $\mathrm{CE}_{\mathrm{val}}$ records the validation loss of the current network and serves to monitor the update rather than to supply a gradient. Equation (1) uses one scalar cost per reference class, whereas Equations (2)–(5) describe a pairwise matrix. The two are linked by reading the cost of class k from the off-diagonal entries of its row (Equation (6)), which expresses the average penalty for misclassifying an example of class k as any other class.

$$w_k = \frac{1}{C-1}\sum_{j \neq k} \xi_{kj} \tag{6}$$

In the reported configuration, classifier parameters are updated after each batch and costs after each epoch. For $C = 14$ labels, the cost matrix is initialized with diagonal entries of 1 and off-diagonal entries of 30, the cost learning rate is 0.1, and $\gamma = 3$. These settings identify the configuration compared with cross entropy, focal loss, and other cost-sensitive alternatives in Section 4.4, and the sensitivity of the results to γ is reported there as well.

FOCC is assessed here as a training configuration within a feasibility study. Its practical advantage over the automatic cost optimization of Khan et al. (2018) lies mainly in computation: separability is obtained from predicted probabilities rather than from pairwise intra- and inter-class distances, and in the reported runs the mean computing time was 8371.1 ± 72.5 s for FOCC compared with 77,708.2 ± 258.9 s for automatic cost optimization. Its contribution to classification performance, by contrast, is small relative to fold variation (Section 4.4), so the formulation above is offered to make the configuration reproducible rather than to claim a performance mechanism.

3.6. Evaluation Design and Implementation

The evaluation uses five folds over the 4904 agreed utterances, with 5% of each training portion held out as the validation set required by the cost update. The folds were formed over utterances, so utterances from the same lecture can appear in both training and test portions. The resulting estimates therefore describe within-corpus classification, and Section 4.3 reports an additional analysis that keeps contiguous stretches of the corpus together to examine how much these estimates depend on shared lecture context.

All neural models were trained with Adam (D. P. Kingma & Ba, 2014) for a fixed schedule of 20 epochs, with batch size 8, learning rate 5×10^{-5}, and random seed 42. Implementation used PyTorch (Paszke et al., 2017) on an NVIDIA RTX 3090 GPU and an Intel i5-12400 CPU. Random Forest and Logistic Regression are conventional comparison methods and fall outside this neural optimization schedule. Table 4 collects these settings.

Macro F1 is the primary metric because it gives each category equal weight; accuracy is reported for completeness but is dominated by the neutral class. Results are presented as means and standard deviations across the five folds. To make the uncertainty of these means explicit, 95% confidence intervals are computed from the fold means and standard deviations with a t distribution on four degrees of freedom. Because fold-level records for paired comparisons are not part of the summarized results, these intervals are used to judge whether differences are large relative to fold variation, not as formal paired tests.

Two imbalance-determined references calibrate the results. An always-neutral rule has accuracy $q = 4180/4904 = 0.852$, neutral F1 $2q/(1 + q)$, and zero F1 for the other 13 labels, giving macro F1 = $[2q/(1 + q)]/14 = 6.57\%$. Random guessing in proportion to the class distribution has an expected F1 for each class equal to its prevalence, giving an expected macro F1 of $1/14 = 7.14\%$ and an expected accuracy of 73.0%. These values indicate what the class distribution alone produces, and they serve as the reference against which macro F1 is interpreted.

The baselines include Random Forest (Breiman, 2001), Logistic Regression, TextCNN, TextRNN, TextRNN with attention, FastText (Joulin et al., 2017), DPCNN (Johnson & Zhang, 2017), BERT (Devlin et al., 2018), BERT–BiLSTM, and BERT–CNN; BiLSTM follows Hochreiter and Schmidhuber (1997). ERNIE (Sun et al., 2019) and RoBERTa (Liu et al., 2019) were also run, with aggregate macro F1 values of 23.7 ± 5.7% and 29.6 ± 5.7%. Their classwise results, however, showed a precision of 0.84 for tch-tk, 0.42 for ddl in the case of RoBERTa, and 0.00 for all remaining labels, which cannot be reconciled with these aggregate values. Because the two sets of figures cannot both describe the same predictions, ERNIE and RoBERTa are not included in the comparative tables, and the 0.6-point difference between the proposed configuration and RoBERTa is not interpreted.

Small categories make precision unstable. For each label, the expected number of test examples per fold is one fifth of its support, and the expected number of validation examples per fold is 4% of its support; both are listed with the classwise results in Section 4.1. Praise has on average 2.8 test examples per fold and ask-stu, ex-ans, and offer-e between 3.4 and 3.6; the validation portion contains on average fewer than one example of each of these categories. With so few positives, a single correct positive prediction yields a precision of 1.00 and a single incorrect one yields 0.00. Classwise precision is therefore reported together with support, and recall and F1 are examined through the reference analysis described next.

To provide category-level recall, F1, and confusion counts on the same consensus corpus, a transparent lexical reference analysis was conducted. Each utterance was represented by TF-IDF weighted character 1–3 g (minimum document frequency 2, sublinear term frequency), and a multinomial logistic regression with class weights inversely proportional to class frequency was trained in each fold (Python 3.13, scikit-

learn 1.9.1). Two partitions were used: stratified five-fold cross-validation with seed 42, which corresponds to the utterance-level design of the main experiments, and five contiguous blocks of the corpus file (utterances 1–981, 982–1962, 1963–2943, 2944–3924, and 3925–4904), which keep neighboring discourse together in either the training or the test portion. Because neighboring utterances largely originate from the same lecture, the contiguous partition approximates the effect of reducing shared lecture context, although it is not a strictly lecture-disjoint split. This analysis is a reference for the structure of the task, not a competitor to the neural configurations, which were trained under a different protocol.

3.7. Corpus-Level Cue Analysis and Input Masking Illustrations

RQ3 is addressed first through a corpus-level cue analysis of all 4904 agreed utterances. Character n-grams were ranked for each category by a smoothed log-odds ratio that compares their frequency in the category with their frequency in the rest of the corpus, which identifies wording that is distinctive of a category rather than merely frequent. For selected cues with a clear relation to a codebook definition, string matching in the agreed ASR text was then used to count the utterances containing the cue, the share of these utterances that carry the target label, the share of the target category that contains the cue, and the other labels that occur among the cue utterances; the same string matching yields the marker percentages in Table 3. These counts are descriptive properties of the consensus corpus. They characterize the relation between wording and coded function in the material, not the decision rule of any classifier, and the results are reported in Section 4.5.

The masking analysis complements the cue analysis by illustrating representation sensitivity for three selected utterances of informational praise, a deadline statement, and neutral teacher talk. The three utterances were selected as typical instances of these behaviors; whether the models classified them correctly is not recorded, so the illustrations carry no information about classification accuracy. For each example, one character is replaced with [MASK], and the L1 change in the final representation is used as an indicator of sensitivity. Equation (7) separates the raw representation change r from its normalized display score a.

$$r(\omega, s) = \sum_{j=1}^{d} |h_s^j - \tilde{h}_{s,\omega}^j|, \qquad a(\omega, s) = \frac{r(\omega,s)}{\max_{v \in s} r(v,s)} \tag{7}$$

Here h denotes the representation of the original segment and $\tilde{h}$ the representation after masking character ω. The raw change sums absolute differences over d representation dimensions. Maximum normalization gives a relative score within the example; a zero score is defined for the all-zero case to complete the written definition. This formalization clarifies the described procedure, and the resulting scores are displayed in Section 4.5.

Representation sensitivity indicates which perturbations alter an internal representation; it is not a measure of the contribution of a character to a correct class decision, and the masking display is therefore read as an illustration rather than as evidence about classification accuracy. Because three examples cannot support claims about the population of utterances, RQ3 is addressed primarily through the corpus-level cue analysis in Section 4.5, which uses all 4904 agreed utterances, with the masking display as a complementary illustration.

4. Results

4.1. Classification Relative to Imbalance-Determined References

Table 5 presents the aggregate results. The BERT–BiLSTM–CNN configuration with FOCC reached a mean macro F1 of 30.2% (95% CI 22.3–38.1). The lower bound of this interval is more than three times the always-neutral reference of 6.57% and the prior-matched reference of 7.14%, and the intervals of all BERT-based configurations lie well above those of Random Forest and Logistic Regression, whose macro F1 values (8.1% and 7.9%) remain close to the references. This contrast shows that the contextual models identify non-neutral categories to an extent that the class distribution alone cannot produce, which is the central feasibility question of RQ2. Accuracy behaves differently: the mean of 86.7% (95% CI 82.0–91.4, computed from Table 5 in the same way as for macro F1) is close to the majority-class rate of 85.2%, so accuracy carries little information about the categories of interest and is not used as evidence of performance.

Table 5. Classification results: five-fold means, fold standard deviations, and 95% confidence intervals for macro F1.

Model	Accuracy (%)	Macro F1 (%)	95% CI for Macro F1
Always-neutral reference	85.2	6.6	–
Prior-matched random reference (expected)	73.0	7.1	–
Random Forest	85.9 ± 0.2	8.1 ± 0.6	7.4–8.8
Logistic Regression	86.1 ± 0.3	7.9 ± 0.6	7.2–8.6
TextCNN	84.5 ± 4.4	13.7 ± 5.3	7.1–20.3
TextRNN	85.2 ± 3.7	15.6 ± 6.2	7.9–23.3
TextRNN with attention	85.5 ± 5.6	16.4 ± 3.8	11.7–21.1
FastText	86.1 ± 6.5	21.6 ± 4.3	16.3–26.9
DPCNN	85.6 ± 4.8	12.8 ± 4.8	6.8–18.8
BERT	80.4 ± 9.2	28.2 ± 4.4	22.7–33.7
BERT–BiLSTM	84.2 ± 6.0	25.6 ± 5.9	18.3–32.9
BERT–CNN	85.6 ± 3.4	29.2 ± 6.2	21.5–36.9
BERT–BiLSTM–CNN with FOCC	86.7 ± 3.8	30.2 ± 6.4	22.3–38.1

Table 6 shows classwise precision for the proposed configuration and for BERT and BERT–CNN. Precision values of 1.00 occur mainly in categories with small support: 13 of the 15 such values in Table 6 fall in categories with fewer than ten expected test examples per fold, so they indicate that the few positive predictions made were correct rather than that the category was reliably detected. The three configurations agree in assigning relatively stable precision to the larger categories, tch-tk (0.89 in each case) and hints (0.53–0.58), and they diverge sharply for small categories such as mk-sud and offer-e, where precision ranges from 0.00 to 1.00 and from 0.20 to 1.00, respectively, across models. This pattern is the expected consequence of small supports, and it motivates the recall and confusion analysis in Section 4.2. For practical use, high precision in a small category means that retrieved examples are likely to be relevant, not that most instances of the behavior are found.

Table 6. Classwise precision of three configurations with corpus support and expected support per fold.

Label	Corpus *n*	Expected Test *n* per Fold	Expected Validation *n* per Fold	BERT	BERT–CNN	BERT–BiLSTM–CNN with FOCC
tch-tk	4180	836.0	167.2	0.89	0.89	0.89
ut-ans	67	13.4	2.7	0.13	0.40	0.29

ut-comd	110	22.0	4.4	0.40	1.00	1.00
mk-sud	29	5.8	1.2	1.00	0.00	0.00
stu-tk	42	8.4	1.7	0.50	0.50	1.00
ddl	64	12.8	2.6	0.64	0.57	0.60
praise	14	2.8	0.6	0.75	1.00	1.00
ask-stu	17	3.4	0.7	0.33	1.00	1.00
comm	35	7.0	1.4	0.08	1.00	1.00
hints	245	49.0	9.8	0.53	0.56	0.58
ex-ans	17	3.4	0.7	1.00	1.00	1.00
con-q	40	8.0	1.6	0.18	0.00	1.00
offer-e	18	3.6	0.7	1.00	0.25	0.20
resp	26	5.2	1.0	0.50	0.70	0.80

4.2. Recall, F1, and Confusion Patterns in the Reference Analysis

Table 7 reports the lexical reference analysis. Under the stratified partition, the transparent model reached a macro F1 of 33.8 ± 2.2% and an accuracy of 82.2 ± 1.4%. Its macro F1 lies in the same range as the neural configurations in Table 5, which reinforces the interpretation that the present evidence concerns the feasibility of the task rather than the superiority of a particular architecture. Recall was substantial for several theoretically central behaviors: 0.86 for hints, 0.69 for responsiveness, 0.67 for deadline statements and encouragement, and 0.52 for invitations to student talk. Recall was low for should statements, perspective taking, controlling questions, offering choice, and exhibiting solutions, which are categories whose instructional function depends most on the surrounding interaction (Section 4.6).

Table 7. Classwise precision (P), recall (R), and F1 of the transparent lexical reference analysis under stratified and contiguous-block partitions.

Label	*n*	Stratified P	Stratified R	Stratified F1	Contiguous P	Contiguous R	Contiguous F1
tch-tk	4180	0.93	0.88	0.90	0.91	0.88	0.90
ut-ans	67	0.13	0.30	0.18	0.04	0.07	0.05
ut-comd	110	0.17	0.21	0.19	0.14	0.19	0.16
mk-sud	29	0.07	0.03	0.05	0.24	0.14	0.17
stu-tk	42	0.26	0.52	0.35	0.24	0.43	0.31
ddl	64	0.45	0.67	0.54	0.45	0.58	0.50
praise	14	0.67	0.29	0.40	0.00	0.00	0.00
ask-stu	17	0.00	0.00	0.00	0.00	0.00	0.00
comm	35	0.12	0.03	0.05	0.00	0.00	0.00
hints	245	0.57	0.86	0.69	0.56	0.78	0.65
ex-ans	17	0.00	0.00	0.00	0.00	0.00	0.00
con-q	40	0.07	0.02	0.04	0.00	0.00	0.00
offer-e	18	0.92	0.67	0.77	0.00	0.00	0.00
resp	26	0.58	0.69	0.63	0.00	0.00	0.00
Macro F1 (%)				33.8 ± 2.2			19.8 ± 3.2
Accuracy (%)				82.2 ± 1.4			80.8 ± 3.0
Orientation macro F1 (%)				60.4 ± 1.4			54.8 ± 2.2

Note. Classwise values are computed from predictions pooled across the five test folds; summary rows give fold means ± standard deviations. Orientation macro F1 collapses the 14 labels into neutral, support, and control.

The confusion counts (Figure 3) show where the difficulty lies. Of the 724 non-neutral utterances, 426 (58.8%) received a non-neutral label and 298 (41.2%) were assigned to neutral teacher talk. Only 36 (5.0%) were assigned to a category of the opposite

orientation: 18 of the 397 supportive utterances (4.5%) received a controlling label, and 18 of the 327 controlling utterances (5.5%) received a supportive label. When the 14 labels are collapsed to the three orientations, macro F1 rises to 60.4 ± 1.4%, with F1 values of 0.90 for neutral, 0.62 for support, and 0.30 for control. The main source of error is thus the boundary between specific indicators and residual teacher talk, rather than confusion between autonomy-supportive and controlling language. This distinction matters for applications: omissions lead to undercounting of behaviors, whereas cross-orientation errors, which would most directly mislead a teacher about the motivational character of their talk, are comparatively rare.

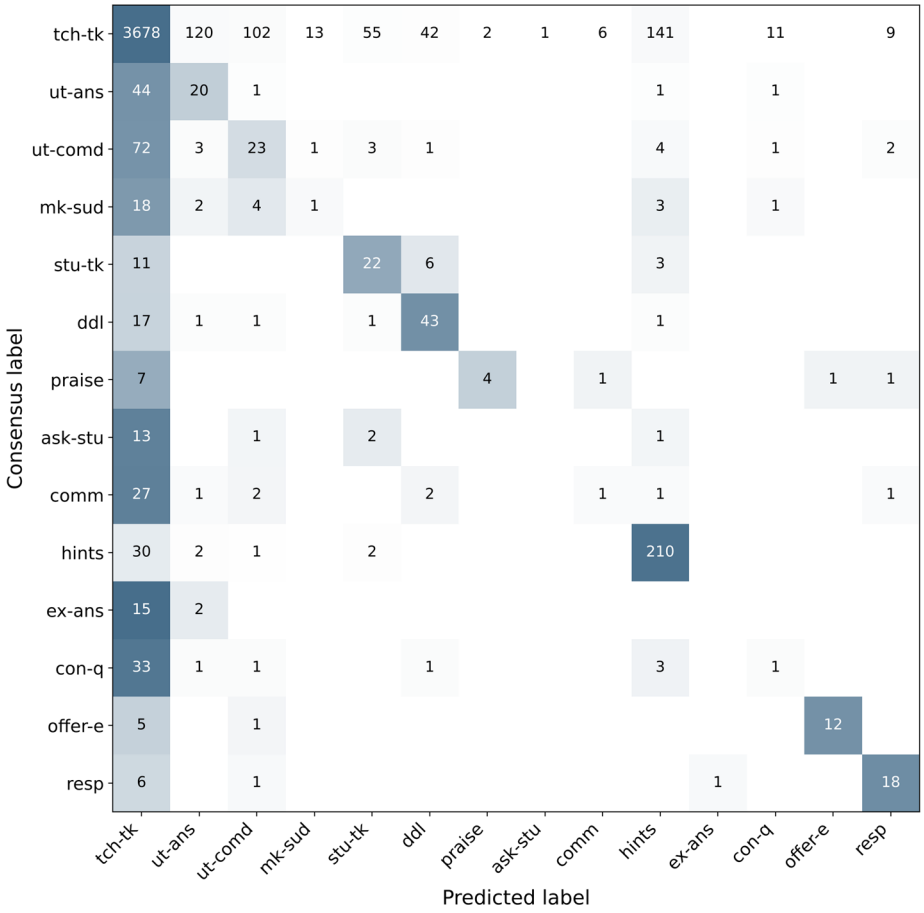


Figure 3. Confusion matrix of the lexical reference analysis under the stratified partition, pooled over the five test folds. Rows are consensus labels and columns are predicted labels; cell shading shows the row proportion.

Figure 3 presents the full confusion matrix. Its first column shows that omissions into neutral talk dominate the errors for every category, and its first row shows the cost of class weighting: 502 neutral utterances (12.0%) received a non-neutral label, most often hints (141), ut-ans (120), and ut-comd (102). This explains why the reference model's accuracy (82.2%) is lower than that of the neural configurations (86.7%) although its macro F1 is similar or slightly higher; the two analyses occupy different points on the precision–recall trade-off. For the intended use, expert-assisted retrieval, the weighting can be adjusted to the purpose: favoring recall when the goal is to collect candidate examples for expert review, and favoring precision when retrieved examples are shown directly to teachers. Outside the neutral row and column, the matrix is sparse: no off-diagonal cell exceeds six utterances, and the 71 utterances in these cells are split almost evenly between pairs within

the same orientation (35), such as mk-sud predicted as ut-comd, and pairs across orientations (36), the largest being stu-tk predicted as ddl (6).

4.3. Sensitivity to Dependence Within the Corpus

When contiguous blocks were held out, macro F1 fell from 33.8% to 19.8 ± 3.2%, whereas accuracy (80.8%) and orientation-level macro F1 (54.8%) changed less (Table 7). Table 8 shows why. Several categories are concentrated in particular stretches of the corpus: 24 of the 26 responsiveness utterances, 16 of the 18 encouragement utterances, and 10 of the 14 praise utterances fall in the fifth segment, and 45 of the 67 ut-ans utterances in the first. When such a segment is held out, the training portion contains only a handful of examples of the category, and its recall drops to zero regardless of how well the category is expressed linguistically. The larger categories, whose examples are spread across segments, retained similar performance: hints (F1 0.69 versus 0.65), deadline statements (0.54 versus 0.50), invitations to student talk (0.35 versus 0.31), and neutral talk (0.90 in both partitions).

Table 8. Distribution of non-neutral categories across the five contiguous corpus segments used in the contiguous-block partition.

Label	*n*	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Largest Segment Share (%)
ut-ans	67	45	4	6	4	8	67
ut-comd	110	25	22	15	29	19	26
mk-sud	29	10	4	7	3	5	34
stu-tk	42	5	6	17	7	7	40
ddl	64	9	10	11	15	19	30
praise	14	4	0	0	0	10	71
ask-stu	17	8	1	5	2	1	47
comm	35	6	4	3	10	12	34
hints	245	34	47	103	32	29	42
ex-ans	17	5	2	3	7	0	41
con-q	40	17	8	3	4	8	43
offer-e	18	0	1	1	0	16	89
resp	26	0	0	0	2	24	92

Two conclusions follow. First, the categories with distributed examples provide evidence that their verbal cues transfer across lecture contexts, since their performance was largely preserved when neighboring discourse was kept out of training. Second, the rare categories are concentrated in particular instructional settings, which is itself an informative finding about how autonomy-supportive and controlling behaviors occur in university lectures: some teachers and activities contribute most of the instances of a given behavior. Utterance-level estimates for these categories therefore partly reflect shared lecture context, and their generalization requires a corpus in which each behavior is observed across several teachers. This is the most direct requirement for the lecture- and teacher-disjoint evaluation discussed in Sections 5.5 and 5.6.

4.4. Representation and Loss Comparisons

Contextual BERT representations yielded clearly higher macro F1 than Word2vec (Mikolov et al., 2013) in the representation comparison (Table 9), in which all variants feed a BiLSTM–CNN trained with cross entropy and the Word2vec input is matched at 3072 dimensions. The difference between Word2vec (11.2 ± 1.1%) and every BERT variant (26.8–29.4%) is large relative to fold variation. Differences among the BERT layer combinations, by contrast, span 2.6 points against fold standard deviations of 5.3–6.4

points. Concatenating the final four layers gave the highest mean, 29.4%, which is consistent with the usefulness of combining information from several encoder layers; because concatenation also changes the downstream input width, this result describes the representation configuration as a whole.

Table 9. Representation comparison with cross entropy.

Representation	Accuracy (%)	Macro F1 (%)
Word2vec	83.1 ± 0.7	11.2 ± 1.1
Final BERT layer	85.2 ± 2.6	26.8 ± 5.6
Sum of final two BERT layers	85.3 ± 3.3	27.9 ± 6.3
Sum of final three BERT layers	84.2 ± 3.5	27.1 ± 6.4
Sum of final four BERT layers	85.6 ± 3.3	28.3 ± 5.8
Concatenation of final two layers	86.4 ± 2.7	28.6 ± 6.1
Concatenation of final three layers	86.3 ± 3.1	29.1 ± 5.3
Concatenation of final four layers	86.5 ± 3.6	29.4 ± 6.4

Table 10 compares loss configurations and two data-level alternatives. Resampling performed worst: random undersampling discarded most of the neutral data and reduced accuracy to 20.7%, and random oversampling reduced macro F1 to 21.5%, which indicates that loss-level treatment of imbalance suits this task better than replicating or discarding examples. Among the loss configurations (Figure 4a), FOCC with γ = 3 had the highest mean macro F1, 0.8 points above cross entropy, but all loss-level configurations lie within 1.6 points of one another, against fold standard deviations of 5.5–6.9 points, and their confidence intervals overlap almost completely. Figure 4b shows that FOCC exceeded the cross-entropy mean for γ between 1 and 5 (29.7–30.2%) and fell below it at γ = 6 (28.8%), so the configuration behaves stably over a broad range of the focal parameter. These results support FOCC as a reasonable and computationally efficient configuration for imbalanced discourse data; they do not establish that it outperforms the alternatives, which would require repeated seeds and paired fold comparisons.

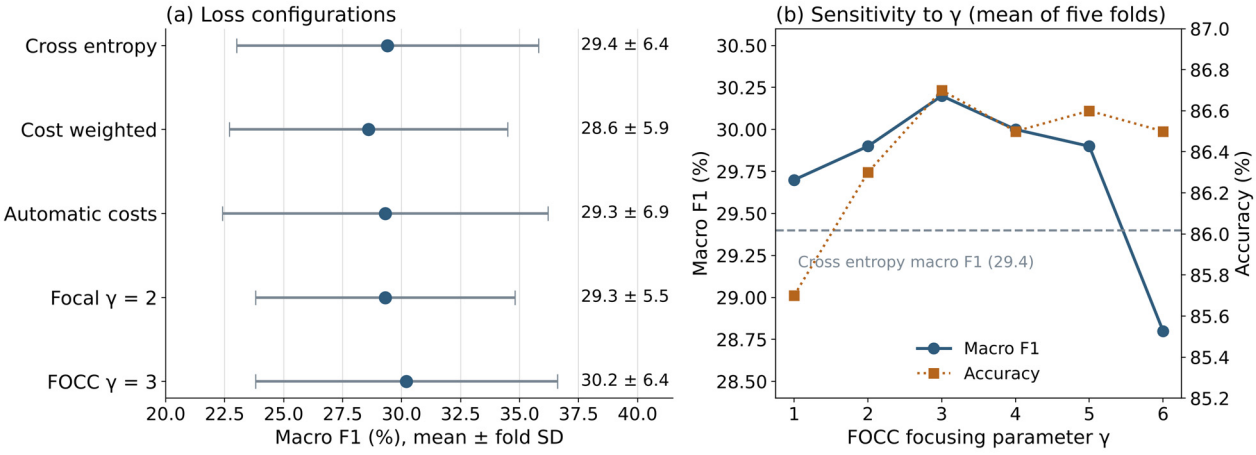


Figure 4. (**a**) Macro F1 across loss configurations; points show five-fold means and bars ± one fold standard deviation. (**b**) Mean macro F1 and accuracy of FOCC for γ = 1–6; the dashed line marks the cross-entropy macro F1.

Table 10. Comparison of loss configurations and data-level resampling.

Configuration	Accuracy (%)	Macro F1 (%)	95% CI for Macro F1
Random undersampling	20.7 ± 3.2	17.2 ± 2.8	13.7–20.7

Random oversampling	86.1 ± 3.5	21.5 ± 5.7	14.4–28.6
Cross entropy	86.5 ± 3.6	29.4 ± 6.4	21.5–37.3
Cost weight sensitive loss	86.2 ± 3.6	28.6 ± 5.9	21.3–35.9
Automatic optimization of class-dependent cost	86.6 ± 3.4	29.3 ± 6.9	20.7–37.9
Focal loss $\gamma = 2$	86.3 ± 3.3	29.3 ± 5.5	22.5–36.1
FOCC $\gamma = 3$	86.7 ± 3.8	30.2 ± 6.4	22.3–38.1

4.5. Corpus-Level Linguistic Cues and Masking Illustrations

Corpus-level cue analysis provides the population-level evidence for RQ3. Distinctive character n-grams, ranked by a smoothed log-odds ratio of each category against the rest of the corpus, correspond closely to the codebook definitions: 比如说 ("for example") occurs in 66% of hints, 应该 ("should") in 90% of should statements, 讨论 ("discuss") in 98% of invitations to student talk, 分钟 ("minute") in 59% of deadline statements, 非常好 ("very good") in 36% of praise, 肯定能 ("surely can") in 33% of encouragement, and 问题 ("question") in 81% of responses to student questions. The categories therefore have recognizable lexical signatures, which is consistent with their identification well above the imbalance-determined references.

Table 11 shows the other side of this relation. The same cues occur far more often outside their target category than inside it: only 7.5% of the 345 utterances containing 应该 were coded as should statements, because the word is also used epistemically ("probably") in ordinary explanation; only 3.4% of utterances containing 问题 were coded as responsiveness; and 35 of the 83 utterances containing 分钟 were coded as neutral, mostly when durations were part of the lecture content. Annotators thus coded instructional function rather than the presence of keywords, which supports the content validity of the labels and indicates that surface cues alone do not determine the coded category. This pattern is consistent with the low recall for should statements and controlling questions in Table 7: their characteristic words are common in neutral talk, and their function is signaled by the demand or pressure they place on students.

Table 11. Selected lexical cues: co-occurrence with their target category and with other categories in the 4904 agreed utterance.

Cue (Gloss)	Target Label	Utterances with Cue	Share Coded as Target (%)	Share of Target Containing Cue (%)	Other Labels Among Cue Utterances
分钟 (minute)	ddl	83	45.8	59.4	tch-tk 35, stu-tk 6
比如 (for example)	hints	325	59.1	78.4	tch-tk 118, stu-tk 4, mk-sud 3
讨论 (discuss)	stu-tk	179	22.9	97.6	tch-tk 108, hints 15, ut-comd 5, ddl 5
应该 (should; probably)	mk-sud	345	7.5	89.7	tch-tk 274, hints 22, ut-comd 7, ut-ans 4, ddl 4
必须 (must)	ut-comd	76	17.1	11.8	tch-tk 56, hints 3
非常好 (very good)	praise	26	19.2	35.7	tch-tk 14, offer-e 3
肯定 (certainly)	offer-e	127	5.5	38.9	tch-tk 100, hints 7, con-q 3, ut-ans 3, ut-comd 3
问题 (question)	resp	624	3.4	80.8	tch-tk 503, hints 31, ut-comd 14, comm 12, stu-tk 10, ddl 8, ut-ans 7, ask-stu 4, con-q 4, mk-sud 3, offer-e 3, ex-ans 3
为什么 (why)	con-q	209	4.8	25.0	tch-tk 157, ut-ans 15, hints 10, ut-comd 7, comm 4, stu-tk 3

Note. Counts are from string matching in the ASR text. Other labels are listed when they account for at least three cue utterances.

The masking illustrations in Figure 5 are consistent with these corpus-level cues. In the praise example, the hybrid configuration's display emphasizes wording translated as "behaved" and "well"; in the deadline example, "left" and "minutes"; and in the neutral example, subject terms such as standard deviation and samples. These displays are illustrations of representation sensitivity for three exemplars, and their value lies in showing that the representation responds to the same kinds of wording that distinguish the categories across the whole corpus. Their instructional significance remains sensitive to context: a time expression may provide structure, and positive evaluation may inform or pressure.

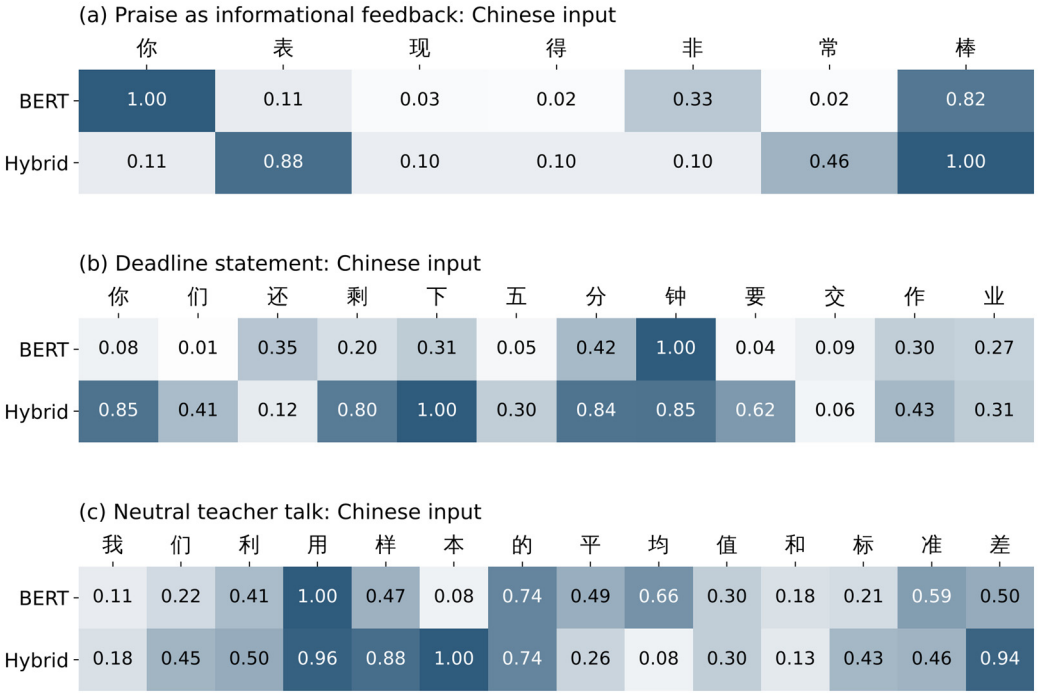


Figure 5. Input masking illustrations for three selected teacher utterances. Chinese characters above the columns represent the original input, with each column corresponding to one character. Rows compare BERT and the hybrid configuration. English translations of the complete utterances are: (**a**) "You behaved very well"; (**b**) "You have five minutes left to hand in your homework"; (**c**) "We use the mean and standard deviation of the samples." Words discussed in the text: 表现 ("behaved") and 棒 ("well") in (**a**); 剩下 ("left") and 分钟 ("minutes") in (**b**); 标准差 ("standard deviation") and 样本 ("samples") in (**c**).

4.6. Codebook Boundaries

Table 12 relates the main codebook boundaries to the evidence observed in the corpus and in the reference analysis. Each boundary concerns a pair of categories that share wording but differ in instructional function, and the corpus evidence shows that these boundaries are where classification errors concentrate. The table therefore links the conceptual definitions to observed model behavior and identifies which distinctions most require context beyond the single utterance.

Table 12. Codebook boundaries with corpus and model evidence.

Boundary	Illustration	Required Interpretive Evidence	Evidence in this Corpus
Hint versus supplied answer	A suggestion about where to start versus giving the completed solution	Whether the learner retains an opportunity to discover; preceding task and attempts	Hints recall 0.86, ut-ans 0.30; 120 neutral utterances predicted as ut-ans
Offering choice versus controlling question	Leaving the way of working to students versus a directive framed as a question	Available choice, preceding dialogue, and response to refusal	ask-stu and con-q recall 0.00 and 0.02; 为什么 ("why") occurs in 209 utterances, of which 10 were coded con-q
Informational praise versus approval pressure	Positive evaluation or "Good job"	Whether feedback concerns progress or is contingent approval; tone and context	非常好 ("very good") occurs in 26 utterances: 5 praise, 3 encouragement, 14 neutral
Deadline versus organization	A reminder of remaining time	Whether the statement imposes pressure or provides needed structure	分钟 ("minute") occurs in 83 utterances: 38 ddl, 35 neutral, 6 stu-tk; ddl recall 0.67
Encouragement versus pressure	"You can do it"	Student difficulty, teacher response, and whether choice is preserved	肯定 ("certainly") occurs in 127 utterances, 7 coded offer-e; offer-e precision 0.92, recall 0.67
Responsiveness versus general reference to questions	Answering a student's question versus mentioning a question in exposition	Whether a preceding student turn exists	问题 ("question") occurs in 624 utterances, 21 coded resp; resp recall 0.69 when segments are shared, 0.00 when held out

5. Discussion

5.1. The Blended Approach and the Research Questions

This exploratory study examined whether an SDT-based blended approach, in which theoretical definitions and independent expert annotation define the target and computational classification extends its identification to authentic classroom speech, can recover verbal indicators of autonomy support and control in Chinese university teacher discourse. Taken together, the results suggest that it can do so to a degree that clearly exceeds what the class distribution alone would produce, and they also delineate the conditions under which this holds. The three research questions are discussed in turn.

With respect to RQ1, the study shows that a subset of the SDT observational repertoire can be translated into utterance-level transcript categories when the adaptation is made explicit. The 14-label scheme retains the functional distinctions of the source framework while documenting where transcript coding departs from observation (Tables 1 and 2), and two psychologists applied it independently with a Cohen's κ of 0.818 on 5203 speech-derived utterances from eight disciplines. This level of agreement indicates that the adapted categories are recognizable to trained judges from teacher language alone. The adaptation also makes differences between observation modalities analytically explicit: a teacher's invitation represents an opportunity for student participation, whereas the duration of student speech concerns what follows it, and a verbal reference to a solution is related to, but distinct from, its demonstration. Such distinctions help researchers select indicators that match the available observation channel and the question being asked.

With respect to RQ2, the evidence can be read at two levels of resolution. At the level of the 14 specific behaviors, the proposed configuration reached a macro F1 of 30.2% (95% CI 22.3–38.1), with a lower bound more than three times the imbalance-determined references of 6.6% and 7.1%, whereas conventional classifiers without contextual representations remained close to those references. At the level of the three orientations

named in RQ2, the reference analysis of the same corpus reached a macro F1 of 60.4%, and only 5.0% of non-neutral utterances were assigned to the opposite orientation. The fine-grained task is thus considerably harder, as would be expected for categories with as few as 14 examples, whereas the broader distinction between autonomy-supportive, controlling, and neutral talk is captured more robustly. The dominant error is the omission of specific indicators into residual teacher talk rather than confusion between supportive and controlling language, a pattern that bears directly on how the outputs can be used (Section 5.4).

With respect to RQ3, the corpus-level cue analysis indicates that the categories carry recognizable lexical signatures, for example 比如说 ("for example") in 66% of hints and 讨论 ("discuss") in 98% of invitations to student talk, which is consistent with their identification above the reference levels. The same analysis shows that these cues are far from sufficient on their own, a point developed in Section 5.2. The masking illustrations in Figure 5 are consistent with these corpus-level patterns; as three exemplars, they are read as illustrations of representation sensitivity rather than as population-level evidence.

The results also clarify the role of the computational components introduced as the third and fourth contributions. The hybrid BERT–BiLSTM–CNN configuration trained with FOCC attained the highest mean macro F1 among the neural configurations compared, but the differences among BERT-based configurations (25.6–30.2%) and among loss-level treatments of imbalance (within 1.6 points) are small relative to fold standard deviations of 4.4–6.9 points, and their confidence intervals overlap substantially. The present data therefore do not support the claim that a particular architecture or loss function is superior. The comparisons that do exceed fold variation concern design choices at a different level: contextual representations outperformed static word embeddings by a wide margin within the same downstream network (26.8–29.4% versus 11.2% macro F1), and loss-level treatment of imbalance was preferable to discarding or replicating examples, which reduced accuracy to 20.7% under undersampling and macro F1 to 21.5% under oversampling. A transparent character n-gram model, although trained under a different protocol, also reached a macro F1 in the same range (33.8%).

These findings suggest that the value of the study lies less in any single component than in the blended pipeline as a whole: a theoretically defined target, consensus expert annotation, contextual representations, imbalance-aware training and evaluation, and transparent analysis of errors. Within this pipeline, the hybrid architecture is one reasonable means of combining contextual, sequential, and local phrase information, and FOCC serves a supporting role. FOCC retains all utterances while increasing the weight of rare categories, performed stably across focal parameters from $\gamma = 1$ to 5 (29.7–30.2%), and is computationally efficient (Section 4.4), which makes it a practical configuration for strongly imbalanced discourse data; whether it yields reliable gains over simpler weighting schemes is a question for designs with repeated seeds and paired fold comparisons (Section 5.6). This framing also has a practical implication. Because the approach does not depend on a specific network, its computational components can be replaced as stronger language models become available, while the theoretical target and the expert reference remain fixed.

5.2. Function over Form: Relating the Findings to SDT and Automated Discourse Research

The cue analysis provides substantive, and not only methodological, evidence about how autonomy support and control are expressed in university teacher talk. The characteristic words of each category occur far more often outside the category than inside it: only 7.5% of the 345 utterances containing 应该 ("should") were coded as should statements (Table 11), because the word is also used epistemically in ordinary explanation, and only 3.4% of utterances containing 问题 ("question") were coded as responsiveness.

Annotators thus coded what an utterance did for students rather than which words it contained. This pattern accords with the SDT view that autonomy support and control are defined by their functional significance rather than their surface form (Reeve, 2009), and with the distinction between structure and control in the circumplex model (Aelterman et al., 2019): a reference to time was coded as a deadline statement in 38 of the 83 utterances containing 分钟 ("minute") and as neutral talk in 35, depending on whether it placed a demand on students' activity. The same evidence indicates where classification is hardest. Should statements and controlling questions, whose characteristic words are common in neutral talk and whose function is signaled by the pressure they convey, are among the categories with the lowest recall, and the codebook boundaries in Table 12 coincide with the concentrations of error. The qualitative and quantitative results therefore converge: the distinctions that require interpretive judgment from annotators are the same distinctions that challenge the classifier.

Relative to earlier automated studies of teacher discourse, the present findings extend rather than replicate their results. Kelly et al. (2018) showed that estimates derived from automatically recognized classroom audio can agree with human coding of question authenticity, and Demszky et al. (2021) showed that a computational measure of uptake is associated with instructional quality. The present study suggests that a comparable route is open for an SDT-defined target, in a different language and at a different educational level, and under markedly stronger class imbalance. It also highlights a consideration that is less prominent in studies of more frequent discourse features: when the behaviors of theoretical interest are rare, imbalance-determined references, category-level recall, and dependence on shared context need to be reported for the results to be interpretable. This reporting practice may be useful for other theory-driven applications in which the target behaviors are similarly infrequent.

5.3. Linguistic Indicators Versus Motivational Effects

The present findings concern agreement with expert-coded verbal indicators. Identifying a theoretically defined linguistic indicator is distinct from establishing its motivational effect. The same utterance, such as "You can do it", can be experienced as encouragement or as pressure depending on the student's situation and the relationship in which it occurs, and a directive can provide welcome structure when students need guidance (Patzak & Zhang, 2025). The labels in this study therefore describe what teachers say in terms of SDT-relevant functions as judged by experts; whether these functions support students' autonomy, competence, and relatedness is an empirical question that the present data do not address, because they include no measures of students' need satisfaction, motivation, engagement, or learning.

The design nonetheless prepares the ground for that question. The labels are anchored in an established observational tradition, the cue analysis supports their content validity, and the classifier provides a means of producing indicator profiles for many lessons at low cost. These properties make it possible to relate verbal indicators to student-level outcomes in subsequent work while preserving the distinction between association and causal effect; Section 5.6 outlines the corresponding designs.

5.4. Implications for Research and Teacher Reflection

For research, the approach offers a way to locate candidate instances of SDT-relevant teacher language across many lessons, so that expert observation can be directed to episodes of interest rather than spread across entire recordings. The error pattern is favorable for this purpose: in the reference analysis, errors are dominated by omissions into neutral talk, while only about one in twenty non-neutral utterances is assigned to the opposite orientation. Retrieved examples are therefore likely to be of the right orientation,

although frequency counts will underestimate how often a behavior occurs, and counts for small categories are unstable. The class weighting can also be adjusted to the purpose, favoring recall when candidates are collected for expert review and precision when examples are shown directly to teachers (Section 4.2).

For teacher reflection, the communication of uncertainty is essential. Automated feedback on classroom discourse has been used to support teachers' reflection (Jacobs et al., 2022), and a study of the Teacher Talk Tool found that such feedback stimulated reflection while evidence of subsequent change in practice was more limited (Kelly et al., 2025). A reflection interface should therefore present each candidate utterance with its surrounding turns, a calibrated confidence score, and the option for the user to correct the label or abstain. Low-confidence predictions and categories with small supports should be flagged as such, and aggregate profiles should be accompanied by an indication of their expected undercount. Above all, automated labels should prompt professional discussion of specific episodes, verified by the teacher or an expert, and should not be used to draw evaluative conclusions about a teacher's practice or for summative appraisal. Given the macro F1 obtained here, this example-based, human-verified workflow is the use that the current evidence supports most directly.

5.5. Limitations

As an exploratory feasibility study, the present work has boundaries that qualify its conclusions and define the agenda for subsequent research. They concern the sample and educational context, the evaluation design, the input channel, and the annotation and category structure.

First, the corpus represents eight Chinese university lectures delivered by seven male teachers across eight disciplines. Disciplinary breadth provides linguistic variation. Because participation depended on the practical conditions of recruiting teachers in authentic classroom settings, the teachers who ultimately took part were all male, and their teaching experience and class sizes reflect this naturally occurring sample rather than a systematic sampling design. The findings therefore cannot yet be assumed to extend to female teachers or to teachers and classes with other profiles, and some rare behaviors are contributed mainly by particular teachers or activities. Educational level and culture are also likely to shape both the frequency and the form of the indicators. University lectures are dominated by extended teacher exposition, which accounts for the large share of neutral talk and the rarity of behaviors such as offering students a choice in how to work. In primary and secondary classrooms, student participation and teacher guidance follow different interaction patterns, with more frequent directives, praise, and management talk, and the boundary between supportive structure and control may be drawn differently. Cultural norms of teacher authority and indirectness in Chinese classrooms may also affect how autonomy support is expressed verbally, for example through suggestions framed as examples, which the prominence of hints in this corpus illustrates. The framework is therefore most directly applicable to lecture-based higher education in Chinese, and its categories and codebook boundaries should be re-examined before it is applied at other educational levels, in other languages, or in seminar-based and online formats.

Second, the main experiments used utterance-level folds, so utterances from the same lecture could appear in both training and test portions. The contiguous-block analysis indicates how much this matters. Macro F1 fell from 33.8% to 19.8% when contiguous segments were held out, whereas orientation-level macro F1 changed less (60.4% versus 54.8%), and categories with examples distributed across the corpus, such as hints (F1 0.69 versus 0.65), deadline statements (0.54 versus 0.50), and invitations to student talk (0.35 versus 0.31), retained similar performance. The decline is thus concentrated in rare

categories whose examples cluster in particular segments, which suggests that the verbal cues of the more widely observed behaviors transfer across lecture contexts, whereas estimates for the rarest behaviors partly reflect shared context. Because the contiguous partition approximates rather than reproduces a lecture-disjoint split, generalization across teachers remains to be established.

Third, transcript-based analysis captures one channel of classroom interaction, and it inherits the errors of automatic speech recognition. Some categories depend on short function words: every should statement contains a modal marker, and 88.1% of ut-ans and 87.5% of con-q utterances contain a question particle or question mark, while praise, encouragement, and responsiveness are the shortest categories, with means of 26.0–28.1 characters against 54.7 for neutral talk (Table 3). A single misrecognized word can therefore alter the apparent function of these utterances, for example by turning a suggestion into an apparent directive. Retaining ASR output keeps the task close to realistic recording workflows, but this realism comes at some cost to the linguistic evidence available for these categories. In addition, several categories, including hints versus supplied answers, invitations versus controlling questions, and responsiveness, depend on surrounding dialogue, and these are among the categories with the lowest recall or the strongest dependence on shared segments (Tables 7, 8 and 12). Because recording was restricted to the teacher's speech by design, so that no student speech or information was collected, categories defined partly by students' contributions, such as responsiveness to student questions, are identified from the teacher's side of the exchange only. Prosodic and visual evidence, which transcripts do not contain, would further help to distinguish encouragement from pressure and exhibiting a solution from supplying an answer.

Fourth, the modeling corpus consists of the 4904 utterances on which both annotators agreed. Re-including the 299 excluded cases would change the share of neutral talk only within 80.3–86.1%, so the imbalance profile is well represented, but the excluded cases are likely to be enriched for boundary cases, and the present estimates therefore describe performance on the clearer examples. Small categories impose a further constraint. With on average fewer than four test examples per fold for categories such as praise, ask-stu, ex-ans, and offer-e, a precision of 1.00 indicates that the few positive predictions were correct rather than that the behavior was reliably detected; recall and confusion counts (Section 4.2) are the more informative indicators of detection for these categories. Finally, some categories involve construct boundaries that depend on context. Deadline statements, for example, can serve organizational rather than controlling functions. The codebook addressed this by coding time references as ddl only when they imposed a demand on students' activity, and the corpus shows that annotators applied this distinction consistently, but a single utterance does not always contain sufficient information to make it with certainty.

5.6. Future Directions

Each of these boundaries points to a concrete next step. The first concerns corpus expansion and group-disjoint evaluation. A larger corpus with lecture and teacher identifiers, sampled to include both female and male teachers and a broader range of teaching experience, class sizes, disciplines, and educational levels, would allow lecture- and teacher-disjoint evaluation in which each behavior is observed across several teachers. Such evaluation should combine macro F1 with per-label precision, recall, and confusion counts for the neural configurations, using repeated seeds and paired estimates, with a validation portion large enough to contain examples of every category. Paired comparisons of this kind would also show whether FOCC yields reliable gains over simpler weighting schemes, and a traceable analysis package of corpus labels, prediction

records, and checkpoint versions would support implementation-level replication, including comparison of alternative matrix-to-weight mappings for FOCC.

The second concerns richer input. Comparing ASR output with manually corrected transcripts for the categories most exposed to recognition errors would quantify the trade-off identified above. Inputs that include preceding and following turns are a natural extension for context-dependent categories, and a coarser support/control/neutral resolution, for which the reference analysis already reaches a macro F1 of 60.4%, may be appropriate for applications that do not require the specific behavior. Beyond these text-based extensions, the dataset could be enriched with multimodal information, such as prosodic features of the teacher's voice and visual cues from classroom video, which could serve as additional model inputs alongside the transcript. Such multimodal input may be particularly informative for distinctions that wording alone does not settle, for example between encouragement and pressure, or between exhibiting a solution and supplying an answer.

The third concerns adjudication and boundary refinement. Adjudicating the 299 excluded cases and reporting category-specific agreement would clarify selection effects, and evaluating models with the adjudicated cases included would show how much performance depends on the clearer examples. Targeted annotation of boundary cases, such as organizational versus pressuring time references, would sharpen the codebook where errors are concentrated, and broader masking samples with explicit word alignment would test whether the illustrative patterns in Figure 5 recur across behaviors.

The fourth concerns the link between verbal indicators and students' experience. Human-verified indicators can be aggregated to the lesson or class level and related to student reports of perceived autonomy support and need satisfaction, using multilevel models that account for students nested in classes and repeated lessons. Comparison with established observer ratings of the same lessons would provide criterion evidence for the indicators, and longitudinal or intervention designs would be required to examine whether changes in teacher language precede changes in student motivation. This sequence would connect computational identification, measurement agreement, and the motivational processes proposed by SDT.

The fifth concerns the extension of the blended approach to large language models. Instruction-tuned LLMs offer a natural route for validating and extending the approach, and the present study provides the components that such work requires: an SDT-informed codebook, a consensus-labeled corpus, imbalance-determined references, and category-level error profiles. Three lines of work appear feasible. First, zero-shot and few-shot prompting, with the codebook definitions supplied as instructions, can test how far general-purpose models recover the 14 behaviors and the three orientations under the comparison protocol set out in Section 2.6. Second, parameter-efficient fine-tuning of an LLM on the consensus corpus, combined with class-dependent weighting of the kind that FOCC provides, would show whether larger pretrained models improve recall for the rare and context-dependent categories, particularly because such models can process the surrounding turns that several categories require. Third, LLMs could be incorporated into the blended workflow itself, for example by proposing candidate utterances and brief rationales that experts then verify, so that the efficiency of generative models is combined with the expert reference that defines the target. In each case, the reliability concerns reviewed in Section 2.6, namely label consistency across runs, rationales unsupported by the utterance, and biased interpretations, would need to be assessed against the expert labels, and LLM outputs would complement rather than replace expert judgment. Because the blended approach separates the definition of the target from the computational model that identifies it, such a transfer leaves its theoretical core unchanged, which makes LLMs a feasible next step for SDT-informed analysis of teacher discourse.

6. Conclusions

This exploratory feasibility study connected SDT-informed behavioral categories with expert annotation and computational analysis of authentic Chinese university teacher discourse. By integrating natural language processing with traditional expert-based methods, the study extends the identification of teachers' autonomy-supportive behavior and proposes an innovative SDT-based blended approach, whose feasibility the present results demonstrate in exploratory form. The 4904-utterance consensus corpus and explicit adaptation rationale provide a concrete basis for studying the verbal expression of autonomy support and control. Classification reached macro F1 values well above the levels determined by class imbalance; errors were concentrated at the boundary with residual teacher talk rather than between supportive and controlling language, and the expert labels were shown to capture instructional function rather than vocabulary. Differences among architectures and loss functions were small relative to fold variation, and rare categories were concentrated in particular lecture segments, which defines the present limits of generalization. The findings support the use of SDT-informed discourse classification for expert-assisted retrieval of classroom examples, and they identify lecture- and teacher-disjoint evaluation, adjudication of ambiguous cases, and validation against students' experiences as the steps required before measurement claims about autonomy support can be made. Because the approach separates the theoretical target from the computational model that identifies it, it can also be extended to large language models once these are validated against the same expert reference.

Author Contributions: Conceptualization, Y.Z.; methodology, J.L. and H.X.; formal analysis, J.L.; investigation, Y.F. and T.Z.; resources, J.L.; data curation, Y.F. and T.Z.; writing—original draft preparation, Y.Z., J.L.; writing—review and editing, Y.Z., Y.F.; visualization, Y.F.; supervision, H.X.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The General Project of the National Social Science Fund of China [No.23BTY075].

Data Availability Statement: The data supporting the findings of this study are not publicly available due to privacy and ethical restrictions related to classroom recordings. The conditions of participant consent and ethical approval do not permit sharing these data outside the research team.

Conflicts of Interest: Author Jian Li was employed by the company Tencent Technology Company Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflicts of interest.

References

Aelterman, N., Vansteenkiste, M., Haerens, L., Soenens, B., Fontaine, J. R. J., & Reeve, J. (2019). Toward an integrative and fine-grained insight in motivating and demotivating teaching styles: The merits of a circumplex approach. *Journal of Educational Psychology*, *111*(3), 497–521. https://doi.org/10.1037/edu0000293.

Ahmadi, A., Noetel, M., Parker, P., Ryan, R. M., Ntoumanis, N., Reeve, J., Beauchamp, M., Dicke, T., Yeung, A., Ahmadi, M., Bartholomew, K., Chiu, T. K. F., Curran, T., Erturan, G., Flunger, B., Frederick, C., Froiland, J. M., González-Cutre, D., Haerens, L., … Lonsdale, C. (2023). A classification system for teachers' motivational behaviors recommended in self-determination theory interventions. *Journal of Educational Psychology*, *115*(8), 1158–1176. https://doi.org/10.1037/edu0000783.

Ashrafi, I., Mohammad, M., Mauree, A. S., Nijhum, G. M. A., Karim, R., Mohammed, N., & Momen, S. (2020). BANNER: A cost-sensitive contextualized model for Bangla named entity recognition. *IEEE Access*, *8*, 58206–58226. https://doi.org/10.1109/ACCESS.2020.2982427.

Black, A. E., & Deci, E. L. (2000). The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: A self-determination theory perspective. *Science Education*, *84*(6), 740–756. https://doi.org/10.1002/1098-237X(200011)84:6<740::AID-SCE4>3.0.CO;2-3.

Boggiano, A. K., Flink, C., Shields, A., Seelbach, A., & Barrett, M. (1993). Use of techniques promoting students' self-determination: Effects on students' analytic problem-solving skills. *Motivation and Emotion*, *17*(4), 319–336.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Bureau, J. S., Howard, J. L., Chong, J. X. Y., & Guay, F. (2022). Pathways to student motivation: A meta-analysis of antecedents of autonomous and controlled motivations. *Review of Educational Research*, *92*(1), 46–72. https://doi.org/10.3102/00346543211042426.

Chen, B., Vansteenkiste, M., Beyers, W., Boone, L., Deci, E. L., Van der Kaap-Deeder, J., Duriez, B., Lens, W., Matos, L., Mouratidis, A., Ryan, R. M., Sheldon, K. M., Soenens, B., Van Petegem, S., & Verstuyf, J. (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and Emotion*, *39*(2), 216–236. https://doi.org/10.1007/s11031-014-9450-1.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Dai, S., Cao, Z., Wang, W., Pang, L., Xu, J., Ng, S.-K., & Chua, T.-S. (2025). Media source matters more than content: Unveiling political bias in LLM-generated citations. In *Proceedings of the 2025 conference on empirical methods in natural language processing*. Association for Computational Linguistics. Available online: https://aclanthology.org/2025.emnlp-main.872/ (accessed on 21 October 2026).

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum.

Deci, E. L., Schwartz, A. J., Sheinman, L., & Ryan, R. M. (1981). An instrument to assess adults' orientations toward control versus autonomy with children: Reflections on intrinsic motivation and perceived competence. *Journal of Educational Psychology*, *73*(5), 642–650.

Deci, E. L., Spiegel, N. H., Ryan, R. M., Koestner, R., & Kauffman, M. (1982). Effects of performance standards on teaching styles: Behavior of controlling teachers. *Journal of Educational Psychology*, *74*(6), 852–859.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist*, *26*(3–4), 325–346. https://doi.org/10.1080/00461520.1991.9653137.

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2024). Can automated feedback improve teachers' uptake of student ideas? Evidence from a randomized controlled trial in a large-scale online course. *Educational Evaluation and Policy Analysis*, *46*(3), 483–505. https://doi.org/10.3102/01623737231169270.

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (Volume 1: Long Papers, pp. 1638–1653). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.130.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, arXiv:1810.04805.

Flink, C., Boggiano, A. K., & Barrett, M. (1990). Controlling teaching strategies: Undermining children's self-determination and performance. *Journal of Personality and Social Psychology*, *59*(5), 916–924.

Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology*, *52*(5), 890–898.

Haerens, L., Aelterman, N., Van den Berghe, L., De Meyer, J., Soenens, B., & Vansteenkiste, M. (2013). Observing physical education teachers' need-supportive interactions in classroom settings. *Journal of Sport and Exercise Psychology*, *35*(1), 3–17.

Hang, C. N., Yu, P.-D., Tan, C. W., & Chiu, D. M. (2025). Beyond search: Measuring LLM performance for scientific literature discovery. In *2025 IEEE international conference on teaching, assessment, and learning for engineering (TALE)*. IEEE. https://doi.org/10.1109/TALE66047.2025.11346619.

Hardre, P. L., & Reeve, J. (2003). A motivational model of rural students' intentions to persist in, versus drop out of, high school. *Journal of Educational Psychology*, *95*(2), 347–356.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56–64. https://doi.org/10.3102/0013189X12437203.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hou, R., Fütterer, T., Bühler, B., Bozkir, E., Gerjets, P., Trautwein, U., & Kasneci, E. (2024). Automated assessment of encouragement and warmth in classrooms leveraging multimodal emotional features and ChatGPT. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education* (Lecture Notes in Computer Science, Vol. 14829, pp. 60–74). Springer. https://doi.org/10.1007/978-3-031-64302-6_5.

Howard, J. L., Bureau, J. S., Guay, F., Chong, J. X. Y., & Ryan, R. M. (2021). Student motivation and associated outcomes: A meta-analysis from self-determination theory. *Perspectives on Psychological Science*, *16*(6), 1300–1323. https://doi.org/10.1177/1745691620966789.

Howard, J. L., Slemp, G. R., & Wang, X. (2025). Need support and need thwarting: A meta-analysis of autonomy, competence, and relatedness supportive and thwarting behaviors in student populations. *Personality and Social Psychology Bulletin*, *51*(9), 1552–1573. https://doi.org/10.1177/01461672231225364.

Huić, A., Pavlin-Bernardić, N., & Vlahović-Štetić, V. (2024). Investigating the circumplex model of (de)motivating teaching styles in higher education. *Psihologijske Teme*, *33*(2), 417–437. https://doi.org/10.31820/pt.33.2.9.

Ilagan, M., Beigman Klebanov, B., & Mikeska, J. (2024). Automated evaluation of teacher encouragement of student-to-student interactions in a simulated classroom discussion. In *Proceedings of the 19th workshop on innovative use of NLP for building educational applications (BEA 2024)* (pp. 182–198). Association for Computational Linguistics. Available online: https://aclanthology.org/2024.bea-1.16/ (accessed on 20 October 2026).

Jacobs, J., Scornavacco, K., Harty, C., Suresh, A., Lai, V., & Sumner, T. (2022). Promoting rich discussions in mathematics classrooms: Using personalized, automated feedback to support reflection and instructional change. *Teaching and Teacher Education*, *112*, 103631. https://doi.org/10.1016/j.tate.2022.103631.

Jang, H., Reeve, J., & Deci, E. L. (2010). Engaging students in learning activities: It is not autonomy support or structure but autonomy support and structure. *Journal of Educational Psychology*, *102*(3), 588–600. https://doi.org/10.1037/a0019682.

Jeno, L. M., Danielsen, A. G., & Raaheim, A. (2018). A prospective investigation of students' academic achievement and dropout in higher education: A self-determination theory approach. *Educational Psychology*, *38*(9), 1163–1184. https://doi.org/10.1080/01443410.2018.1502412.

Jiang, A. L., & Zhang, L. J. (2021). University teachers' teaching style and their students' agentic engagement in EFL learning in China: A self-determination theory and achievement goal theory integrated perspective. *Frontiers in Psychology*, *12*, 704269. https://doi.org/10.3389/fpsyg.2021.704269.

Johansen, M. O., Eliassen, S., & Jeno, L. M. (2023). The bright and dark side of autonomy: How autonomy support and thwarting relate to student motivation and academic functioning. *Frontiers in Education*, *8*, 1153647. https://doi.org/10.3389/feduc.2023.1153647.

Johnson, R., & Zhang, T. (2017). Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th annual meeting of the association for computational linguistics* (Volume 1: Long Papers, pp. 562–570). Association for Computational Linguistics. Available online: https://aclanthology.org/P17-1052/ (accessed on 19 October 2026).

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics*. Association for Computational Linguistics. Available online: https://aclanthology.org/E17-2068/ (accessed on 18 October 2026).

Kelly, S., Guner, G., Hunkins, N., & D'Mello, S. K. (2025). High school English teachers reflect on their talk: A study of response to automated feedback with the teacher talk tool. *International Journal of Artificial Intelligence in Education*, *35*(2), 879–913. https://doi.org/10.1007/s40593-024-00417-x.

Kelly, S., Olney, A. M., Donnelly, P., Nystrand, M., & D'Mello, S. K. (2018). Automatically measuring question authenticity in real-world classrooms. *Educational Researcher*, *47*(7), 451–464. https://doi.org/10.3102/0013189X18785613.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., & Togneri, R. (2018). Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(8), 3573–3587. https://doi.org/10.1109/TNNLS.2017.2732482.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*, arXiv:1412.6980.

Kingma, T., Smits, A., Jaarsma, D., & Voogt, J. (2024). What need-supportive and need-thwarting teaching behaviors do university teachers use in their honors classes? An observational study. *International Journal of Educational Research Open*, *6*, 100331. https://doi.org/10.1016/j.ijedro.2024.100331.

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, *9*(3), 231–251. https://doi.org/10.1007/s10984-006-9015-7.

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research*, *86*(2), 602–640.

Leenknecht, M. J. M., Wijnia, L., Loyens, S. M. M., & Rikers, R. M. J. P. (2017). Need-supportive teaching in higher education: Configurations of autonomy support, structure, and involvement. *Teaching and Teacher Education*, *68*, 134–142. https://doi.org/10.1016/j.tate.2017.08.020.

Levesque, C., Zuehlke, A. N., Stanek, L. R., & Ryan, R. M. (2004). Autonomy and competence in German and American university students: A comparative study based on self-determination theory. *Journal of Educational Psychology*, *96*(1), 68–84. https://doi.org/10.1037/0022-0663.96.1.68.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988). IEEE. Available online: https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html (accessed on 19 October 2026).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv*, arXiv:1907.11692.

Long, Y., Luo, H., & Zhang, Y. (2024). Evaluating large language models in analysing classroom dialogue. *npj Science of Learning*, *9*, 60. https://doi.org/10.1038/s41539-024-00273-3.

Lozano-Jiménez, J. E., Huéscar, E., & Moreno-Murcia, J. A. (2021). Effects of an autonomy support intervention on the involvement of higher education students. *Sustainability*, *13*(9), 5006. https://doi.org/10.3390/su13095006.

Malik, M., Malik, M. K., Mehmood, K., & Makhdoom, I. (2021). Automatic speech recognition: A survey. *Multimedia Tools and Applications*, *80*(6), 9411–9457.

McLachlan, S., & Hagger, M. S. (2010). Effects of an autonomy-supportive intervention on tutor behaviors in a higher education context. *Teaching and Teacher Education*, *26*(5), 1204–1210. https://doi.org/10.1016/j.tate.2010.01.006.

Melo, C., de la Maza, J., & Recabarren, M. (2026). Validating AI-generated classroom observations: Reliability, accuracy, and limits of LLM-based pedagogical judgment. *Computers & Education: Artificial Intelligence*, *10*, 100612. https://doi.org/10.1016/j.caeai.2026.100612.

Metzner, O., Wang, Y., de Melo, G., Symes, W., Huang, Y., & Lazarides, R. (2026). The potential and limitations of large language models for automatic classification of teachers' motivational messages in educational research. *British Journal of Educational Psychology*, *96*(1), 14–31. https://doi.org/10.1111/bjep.70013.

Metzner, O., Wang, Y., Symes, W., Huang, Y., Keller, L., de Melo, G., & Lazarides, R. (2025). A process-oriented perspective on pre-service teachers' self-efficacy and their motivational messages: Using large language models to classify teachers' speech. *British Journal of Educational Psychology*, *95*(Suppl. S1), S73–S97. https://doi.org/10.1111/bjep.12779.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*, arXiv:1301.3781.

Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, *88*(2), 203–214.

Niemiec, C. P., & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and Research in Education*, *7*(2), 133–144. https://doi.org/10.1177/1477878509104318.

Okada, R. (2023). Effects of perceived autonomy support on academic achievement and motivation among higher education students: A meta-analysis. *Japanese Psychological Research*, *65*(3), 230–242. https://doi.org/10.1111/jpr.12380.

Orsini, C., Binnie, V. I., & Wilson, S. L. (2016). Determinants and outcomes of motivation in health professions education: A systematic review based on self-determination theory. *Journal of Educational Evaluation for Health Professions*, *13*, 19. https://doi.org/10.3352/jeehp.2016.13.19.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. In *NIPS Autodiff workshop*. Available online: https://openreview.net/forum?id=BJJsrmfCZ (accessed on 19 October 2026).

Patall, E. A., Steingut, R. R., Vasquez, A. C., Trimble, S. S., Pituch, K. A., & Freeman, J. L. (2018). Daily autonomy supporting or thwarting and students' motivation and engagement in the high school science classroom. *Journal of Educational Psychology*, *110*(2), 269–288. https://doi.org/10.1037/edu0000214.

Patzak, A., & Zhang, X. (2025). Blending teacher autonomy support and provision of structure in the classroom for optimal motivation: A systematic review and meta-analysis. *Educational Psychology Review*, *37*, 17. https://doi.org/10.1007/s10648-025-09994-2.

Reeve, J. (2009). Why teachers adopt a controlling motivating style toward students and how they can become more autonomy supportive. *Educational Psychologist*, *44*(3), 159–175. https://doi.org/10.1080/00461520903028990.

Reeve, J., Bolt, E., & Cai, Y. (1999). Autonomy-supportive teachers: How they teach and motivate students. *Journal of Educational Psychology*, *91*(3), 537–548.

Reeve, J., & Cheon, S. H. (2021). Autonomy-supportive teaching: Its malleability, benefits, and potential to improve educational practice. *Educational Psychologist*, *56*(1), 54–77. https://doi.org/10.1080/00461520.2020.1862657.

Reeve, J., & Cheon, S. H. (2024). Learning how to become an autonomy-supportive teacher begins with perspective taking: A randomized control trial and model test. *Teaching and Teacher Education*, *148*, 104702. https://doi.org/10.1016/j.tate.2024.104702.

Reeve, J., Deci, E. L., & Ryan, R. M. (2004). Self-determination theory: A dialectical framework for understanding socio-cultural influences on student motivation. In D. M. McInerney, & S. Van Etten (Eds.), *Big theories revisited* (Vol. 4, pp. 31–60). Information Age Press.

Reeve, J., & Jang, H. (2006). What teachers say and do to support students' autonomy during a learning activity. *Journal of Educational Psychology*, *98*(1), 209–218.

Rogers, C. R., & Freiberg, H. J. (1994). *Freedom to learn* (3rd ed.). Merrill.

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Press.

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, *61*, 101860. https://doi.org/10.1016/j.cedpsych.2020.101860.

Seidel, S. B., Reggi, A. L., Schinske, J. N., Burrus, L. W., & Tanner, K. D. (2015). Beyond the biology: A systematic investigation of noncontent instructor talk in an introductory biology course. *CBE—Life Sciences Education*, *14*(4), ar43. https://doi.org/10.1187/cbe.15-03-0049.

Sheldon, K. M., & Krieger, L. S. (2007). Understanding the negative effects of legal education on law students: A longitudinal test of self-determination theory. *Personality and Social Psychology Bulletin*, *33*(6), 883–897. https://doi.org/10.1177/0146167207301014.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (2013). The classroom observation protocol for undergraduate STEM (COPUS): A new instrument to characterize university STEM classroom practices. *CBE—Life Sciences Education*, *12*(4), 618–627. https://doi.org/10.1187/cbe.13-08-0154.

Stains, M., Harshman, J., Barker, M. K., Chasteen, S. V., Cole, R., DeChenne-Peters, S. E., Eagan, M. K., Jr., Esson, J. M., Knight, J. K., Laski, F. A., Levis-Fitzgerald, M., Lee, C. J., Lo, S. M., McDonnell, L. M., McKay, T. A., Michelotti, N., Musgrove, A., Palmer, M. S., Plank, K. M., … Young, A. M. (2018). Anatomy of STEM teaching in North American universities. *Science*, *359*(6383), 1468–1470. https://doi.org/10.1126/science.aap8892.

Su, Y.-L., & Reeve, J. (2011). A meta-analysis of the effectiveness of intervention programs designed to support autonomy. *Educational Psychology Review*, *23*(1), 159–188. https://doi.org/10.1007/s10648-010-9142-7.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., & Wu, H. (2019). ERNIE: Enhanced representation through knowledge integration. *arXiv*, arXiv:1904.09223.

Tan, M., & Demszky, D. (2026). Do as I say: What teachers' language reveals about classroom management practices. *Educational Researcher*, Advanced online publication. https://doi.org/10.3102/0013189X251410178.

Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology*, *72*(5), 1161–1176.

Van den Berghe, L., Soenens, B., Vansteenkiste, M., Aelterman, N., Cardon, G., Tallir, I. B., & Haerens, L. (2013). Observed need-supportive and need-thwarting teaching behavior in physical education: Do teachers' motivational orientations matter? *Psychology of Sport and Exercise*, *14*(5), 650–661.

Vansteenkiste, M., Ryan, R. M., & Soenens, B. (2020). Basic psychological need theory: Advancements, critical themes, and future directions. *Motivation and Emotion*, *44*(1), 1–31. https://doi.org/10.1007/s11031-019-09818-1.

Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, *87*(2), 246–260.

Vermote, B., Aelterman, N., Beyers, W., Aper, L., Buysschaert, F., & Vansteenkiste, M. (2020). The role of teachers' motivation and mindsets in predicting a (de)motivating teaching style in higher education: A circumplex approach. *Motivation and Emotion*, *44*(2), 270–294. https://doi.org/10.1007/s11031-020-09827-5.

Wang, D., Tao, Y., & Chen, G. (2024). Artificial intelligence in classroom discourse: A systematic review of the past decade. *International Journal of Educational Research*, *123*, 102275. https://doi.org/10.1016/j.ijer.2023.102275.

Wang, D., Vogt, R., Mason, M., & Sridharan, S. (2008). Automatic audio segmentation using the generalized likelihood ratio. In *2008 2nd international conference on signal processing and communication systems* (pp. 1–5). IEEE.

Wang, Y., Luo, W., Liao, X., & Zhao, P. (2024). Exploring the effect of teacher autonomy support on Chinese EFL undergraduates' academic English speaking performance through the mediation of basic psychological needs and classroom engagement. *Frontiers in Psychology*, *15*, 1323713. https://doi.org/10.3389/fpsyg.2024.1323713.

Whitehill, J., & LoCasale-Crouch, J. (2024). Automated evaluation of classroom instructional support with LLMs and BoWs: Connecting global predictions to specific feedback. *Journal of Educational Data Mining*, *16*(1), 34–60. https://doi.org/10.5281/zenodo.10974824.

Wijnia, L., Loyens, S. M. M., & Derous, E. (2011). Investigating effects of problem-based versus lecture-based learning environments on student motivation. *Contemporary Educational Psychology*, *36*(2), 101–113. https://doi.org/10.1016/j.cedpsych.2010.11.003.

Williams, G. C., & Deci, E. L. (1996). Internalization of biopsychosocial values by medical students: A test of self-determination theory. *Journal of Personality and Social Psychology*, *70*(4), 767–779.

Zhou, L., Gao, Y., Hu, J., Tu, X., & Zhang, X. (2022). Effects of perceived teacher support on motivation and engagement amongst Chinese college students: Need satisfaction as the mediator. *Frontiers in Psychology*, *13*, 949495. https://doi.org/10.3389/fpsyg.2022.949495.

Zhou, M., Ma, W. J., & Deci, E. L. (2009). The importance of autonomy for rural Chinese children's motivation for learning. *Learning and Individual Differences*, *19*(4), 492–498. https://doi.org/10.1016/j.lindif.2009.05.003.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.